DOCUMENT RESUME

ED 092 590 TM 003 702

AUTHOR Jenkins, W. O.
TITLE Quick and Dirty Statistics Revisited: The Uses and Abuses of Statistical Analyses in Behavioral Research.
PUB DATE 1 Jan 67
NOTE 139p.

EDRS PRICE MF-$0.75 HC-$6.60 PLUS POSTAGE
DESCRIPTORS Analysis of Covariance; Analysis of Variance; *Behavioral Science Research; Correlation; Data Analysis; *Problems; *Research Design; *Statistical Analysis; Statistical Data

ABSTRACT

This paper is an ardent plea for simplifying experimental design and the associated statistics. The emphasis is on design itself. Traditional designs from simple to complex and reviewed and the simplest, most basic ways of handling the data are presented. Design is stressed in such a way that simple statistics follow. The intactness of inspectional analysis is heavily stressed. Assessment of experimental outcomes in terms of both consistency and magnitude measures is considered at length. The necessity of examining the data from all angles is indicated. The basic role of design and the secondary role of statistics is discoursed on at length. (Author)

ED 092590

TM 003 702

Dr. Deutler

TM

QUICK AND DIRTY STATISTICS REVISITED: THE USES ANL ABUSES OF STATISTICAL ANALYSIS IN BEHAVIORAL RESEARCH

BEST COPY AVAILABLE

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

W.O. Jenkins

January 1, 1967

The Institute for Child Development and Experimental Education

of The City University of New York

and

The Center for Urban Education

ABSTRACT

This paper is a follow-up to one written ten years ago in which an ardent plea and pitch was made for simplifying experimental design and the associated statistics. The plea is repeated here long and loud. The emphasis in the earlier paper was on statistics per se. Here it shifts to design itself. Traditional designs from simple to complex and reviewed and the simplest, most basic ways of handling the data are presented. Design is stressed in such a way that simple statistics follow. The Virgo Intacta of inspectional analysis is heavily stressed. Assessment of experimental outcomes in terms of both consistency and magnitude measures is considered at length. The necessity of examining the data from all angles is indicated. The basic role of design and the secondary role of statistics is discoursed on at length.

"I am grateful to Miss Linda Cutler, Miss Susan J. Marks and Dr. W.E. Morris whose careful reading of all or parts of this manuscript improved it."

TABLE OF CONTENTS

TOPIC	PAGE
Background	1
Layout	3
What's Wrong With Traditional Statistics	4
Virgo Intacta: Inspectional Analysis	6
Replication	8
Backsliding, Data Drift, Regression and Chance	9
The Smaller The N The Better	12
The Case of Deviant Cases	13
Nose Counting and The Binomial Expansion	14
Nose Counting, Association and Correlation	18
How Big? Magnitude Considerations	24
Magnitude: The Case of Two Groups	25
1. Independent Groups	25
2. Matched Pair of Self-Control Design	26
3. Matched Groups	31
Magnitude: The Case of Three or More Groups With One Dimension of Experimental Variation: Single Classification Anova	34
1. Rank Anova	35
2. Classical Anova	37
3. The Anova Range Test	40
Magnitude: Two "Simultaneous" Dimensions of Experimental Variation, Double Classification Anova	44
1. Correlated Data	45
2. Repeated Measurements	52
3. "Simple" Factorial Design	59
4. Experimental Examples of Factorial Design	68
5. Factorial Design: More Than Two Treatment Groups	71

TOPIC	PAGE
Magnitude: More Than Two "Simultaneous" Dimensions of Experimental Variation	77
Magnitude: Analysis of Covariance (Ancova): Partialling Out The Effects of One Variable Upon Another	85
Afterthoughts, Odds and Ends and Some Overview Matters in Experimental Design, Methodology and Statistics	101

QUICK AND DIRTY STATISTICS REVISITED: THE USES AND ABUSES OF STATISTICAL ANALYSIS IN BEHAVIORAL RESEARCH

B A C K G R O U N D

About 12 years ago I wrote a paper entitled "Quick and dirty statistics" (Jenkins, 1955). By that time I had become fed up with textbooks in "experimental design" that dealt almost entirely with statistics and had little or nothing to do with design per se. Design is a first order of business and has its own special problems; statistics are a long second and are determined by the design layout. This point seems obvious, but maybe it isn't. In any event, the purpose of the original paper was to provide researching graduate students with shortcut, rough and ready methods of treating data so they could spend minimal time on analysis and maximal time on research - the proper province of behavioral science. The paper was never published; it was too big and bulky, containing too many tables. Furthermore, I didn't feel like going through the nitpicking process of publication either journal or book. An abbreviated edition of it was issued for hospital personnel interested in research under the heading "Shortcut techniques in the treatment of experimental results" (1956).

Another instigator, tying in with the first, was the continuation of a trend I deprecated in another unpublished paper of some 12 years ago, entitled "On the worship of large numbers" (Jenkins, 1955). Large numbers are real, but not divine. If behavioral scientists paid the respect to chance that they pay to large Ns, the field would be farther advanced and, more importantly, fewer papers would clutter up the journals. Part of the mystique (or possible the psychopathology) of the behavioral scientist is his magical

faith that large Ns will somehow accomplish something. They do: more work for E.

A third item triggering off this paper was a brochure recently handed me at a boat show dealing with Tennessee beer tax facts for 1964. It contains interesting data relating beer consumption to tax rate. The statement is made: "States that have the highest total tax rate (state, city and county) on beer generally have the lowest per capita consumption rate". (It's obvious the pitch is for reduced beer taxes in the state of Tennessee.) I have a powerful aversion to the word "generally". "Generally" speaking, the word "generally" is loose, sloppy, vague and misleading.

For these reasons, this paper was written. The original Quick and Dirty manuscript was short on words and long on tables; the present one is long on words and short of tables. It is not immediately obvious which approach changes more behavior.

This paper could have gone under the guise of several other titles: "Statistics and other minor methodological matters"; "Why mess with complicated statistics when simple ones will do?", "Large numbers really don't make that much difference"; "Statistics in proper perspective"; "There is no magic in statistics or large numbers"; "Statistics made simple"; "How not to analyze data"; "Mistakes we make in treating experimental results"; "The making and breaking of the statistical habit"; "Statistics are real, but not divine"; "How to read data"; "Statistics the easy way"; "The complete guide to understanding numbers"; and so forth.

L A Y O U T

There are two basic types of set-ups to determine whether covariation exists between a stimulus dimension and some measure of behavior. The first is the classical experimental one in which variables are manipulated and functional relationships emerge or not as the case may be. The other is correlational in which measurement (but not manipulation) of two variables is taken and the intensity of relationship or association between them determined. The present write-up will consider both.

There are two additional aspects to most data thatrequire consideration. The effect of an experimental treatment can be "whopper", i.e., large differences in magnitude among the several conditions. Or it can be consistent with every S or pair of Ss showing the impact of the treatment. The two indices can be independent, e.g., small magnitude, but high consistency, but in the limiting case they converge, e.g., when magnitude is large, consistency is high. Investigators should always consider both these aspects of their data. Both will be examined in the present conte.

To facilitate communication, it might be helpful to spe.l out the types of experimental designs to be considered in later sections of this paper - not necessarily in the order given below. The design obviously fixes the limits of the class of statistical analysis to be applied after the data are in; the nature of the data, convenience and personal preference determine what specific class members will be employed. The breakdown of the designs follows. It includes most of those commonly used.

A. One Dimension of Experimental Variation

1. Two groups: independent or randomly assigned groups; matched groups; matched pairs or self-control.

2. More than two groups: single classification analysis of variance (anova).

B. Two "Simultaneous" Dimensions of Experimental Variation: The Effects of Two Variables and Their Interaction

1. Anova for correlated data: matched trios or self-control.

2. Repeated measurements: independent groups treated across blocks of trials or time.

3. "Simple" Analysis of Covariance (Ancova): partialling out pre-treatment differences from treatment measures.

4. "Simple" Factorial Design: two experimental treatments applied "simultaneously".

C. More Than Two "Simultaneous" Dimensions of Variation

1. Complex Anova: three or more variables and their interactions.

2. Complex Ancova: correcting differences in treatment measures for differences in two or more initial pre-treatment indices.

WHAT'S WRONG WITH TRADITIONAL STATISTICS?

There are many things wrong about traditional statistics. For one thing they take too long. For another they're difficult to communicate. But the main thing wrong with them is that they lose sight of the behavior of organ-

isms. Statistics are tools to help simplify and clarify behavioral measurement. If they do less than this - and they frequently do - they detract from rather than contribute to the detection of behavioral principles. Tables of sums of squares, degrees of freedom, interactions and the like are dandy and elegant, but they tell nothing about the behavior of individual organisms. As a matter of fact they obscure and confound it. So why use them? Without going psychoanalytic, psychologists seem to possess some blind faith that fancy statistical analysis will produce an emergent from the data, will refine and go beyond them. This is clear nonsense. The fault is not really in the statistics, but in the design and most probably in the problem selected and particularly the corner into which investigators paint themselves by their selection of experimental treatments and behavioral measurements. Be that as it may, it seems to be a case of "Please don't eat the statistical daisies".

There is another way. Problems can be selected and experiments designed so that simple enumerative statistics can be employed. Count statistics are what count - in more ways than one. With small Ns a quick look-see will immediately reveal how many Experimental cases exceed the highest or average Control case. It's a matter of how to analyze data without really trying - or at least without really working at it. Most numbers are simple, but they can be made complicated and even incomprehensible by appropriate statistical manipulation. These's an old Balkan saying: "There are a thousand doors to let out life, but very few to let it in". Similarly, there are many ways of cutting, slicing and working over data, but few of them carry the message of clarifying and simplifying the original numbers representing the behavior of individual organisms.

VIRGO INTACTA: INSPECTIONAL ANALYSIS

While the phraseology may be redundant, "intact virgin" is an accurate description of the state of the art in looking at behavioral results. Of course, it's true that if one is attempting to relate 10 "personality" measures to 10 "perceptual" ones simultaneously, it's not easy to scan the data to see what's going on. Ignoring the limit and considering the straightforward instance, the first step in any treatment of data is visual scanning, a looksee inspection, to determine what the naked eye can find. (Visual "sequential" or "trend" analysis a la Skinnercumulative recordings are, of course, highly desirable.) If this procedure yields little return, then it seems unlikely that any amount of complicated statistical torturing of the data will help. Besides, negative findings are real and basic. To show a variable has little behavioral impact over a wide range may be more important in many instances than teasing out a large - N difference barely attaining the 5% level of significance. Enormous time and effort can be saved by the simple device of inspectional analysis. If half the Ss produce increased behavior and the other half decreased, why analyze further? Or if half a set of correlations of Chi Squares or any other index are positive and the other half negative, isn't this chance finding meaningful in itself? Again, if means differ by a couple of points and ranges amount to several hundred, there is no statistical way of squeezing anything from the data. More importantly, there is no reason to analyze. The numbers descriptive of the behavioral events that occurred stand on their own little feet.

One point that is puzzling in this connection is why drawing conclusions

about experimental agreement with chance isn't worth doing. A real chance finding is, by definition, a rare event and calls for considerable comment. I have been shown a set of 100 Chi Squares, half positive and half negative, (disregarding magnitude) and have been most impressed with chance while the exhibitor of the numbers strongly desired to make something of the handful of large values indicating a positive relationship. Chance is real, but hard to come by. When it occurs in pure form it surely warrants comment.

For purposes of dialogue I am oversimplifying to some extent, but not overly. In many instances a quick and dirty check of the data answers the question asked. On other occasions, of course, manipulation must be resorted to - minor in a number of cases. Below are given a small set of numbers that superficially resembles nothing more than a hodgepodge:

X	Y
27	61
17	73
30	81
19	52
20	56
35	73
13	76

It is by no means obvious what has happened in these numbers. Arranging the X column in order of magnitude or, even better, plotting both sets of numbers graphically, immediately clears the air. From a graphical representation it is immediately obvious that a curvilinear, U-shaped relationship

has emerged with high and low values of X going with high values of Y and middle values of X going with low values of Y. Thus as X increases, values on Y first decrease, then increase - a straightforward proposition. A few advanced graduate students have failed this type of item on their doctoral written examinations because they failed to see the obvious. Experimentally it has been demonstrated that the same information is communicated many times more rapidly and accurately in graphical than tabular form. If quick visual check does not provide the immediate answer as to what's happened, transformation of numbers to graphical representation will.

How many behavioral scientists visually cut and slice their data before feeding it into some sort of machine? Many apparently do not look. The aversion toward numbers stamped in by grammar school harridans teaching arithmetic may well generalize. The safe way is let the machine do it. But the machine knows nothing of the flaws and foibles of behavior - other than those of its programmer who feeds it. This is neither a plea for nitpicking nor an anti-machine polemic - although there is a place under the sun for both. It is an appeal to behavioral investigators to so select their problems and design their experiments that they can get immediate feedback from the behavioral data, i.e., see immediately what, if anything, happened.

REPLICATION

Psychologists who are supposedly statistically sophisticated exhibit a surprising naivete about chance. In the limiting case a coin will stand on edge if one flips it enough times. Short of that but still extreme, we all are aware that one S's response on one occasion does not make a behav-

ioral phenomenon - unless it happens to be a record mile run or pole vault. What we fail to recognize is that chance is real and five times in 100 will produce findings significant at the 5% level. The only antidote to chance is replication. Only if the same direction of effect holds up on two or (preferrably) more occasions can we start to buy the phenomenon, i.e., bet heavily that the same direction will turn up on the next experimental occasion. Chance will on a very few, fortunately rare, occasions produce an inverted generalization decrement function or greater resistance to extinction after 100% rather than partial reinforcement. What we are betting on, however, is the bulk of the instances, the preponderance of the evidence. "Replication" with variation adds generality to the effect and relieves boredom for E. If one wishes to maximize chance, don't replicate and draw conclusions; if one wishes to minimize chance, replicate before drawing conclusions so that "data drift" is forestalled or at least uncovered.

BACKSLIDING, DATA DRIFT, REGRESSION AND CHANCE

One classic example of backsliding is an investigation during W.W. II where a number of physiological measures were applied to a small sample of pilot trainees. One hundred measures were used on 20 pilots and correlations were computed against the criterion of pass-fail in flight training. By judicious selection, the investigators were able to cull out three measures (of the 100) that generated a multiple correlation of about .98. They drew sweeping conclusions. They were asked, of course, to replicate and they did, reporting another multiple R of .97. It was, of course, based on

three entirely different measures from those of the first study. When the original three measures were employed with the second sample, the multiple correlation naturally became .00. Clearly, their measures were useless in this context. This is a beautiful case of the operation of chance. The investigators so stacked the chance cards against themselves that they couldn't possibly win. One should give chance a chance, but not maximize its operation.

Likewise, as a tour de force, I once analyzed the boxscores of 10 baseball games to determine the relationship between winning or losing and number of players employed on the expectation that the losing team throws in more players. The first time I did it the Phi Coefficient came out around .90. This looked too good so I took 10 consecutive sets of 10 ball games each and applied the same procedure. The resulting Phis were: .50, .30, .50, .61, .40, .31, .30, .20, and .73. The average of these is a shade above .40, considerably less than half of the original correlation. Again, instances could be multiplied, but the point is clear: chance is real.

Again, in another context I have related the amount of money raised to the number of reported cases of various diseases and disorders such as cancer and polio for the year 1958. The numbers are confusing, but fascinating. Correlational procedures applied to them yield a Phi of -.25 based on a cut-off at the means, one of plus .20 with a cut-off at the medians and a rank order Rho of -.43. Further sets of figures and replication are clearly needed.

I was once presented with a mass of t-ratios relating personality meas-

ures to perceptual test performance. Overall, there were 93 positive and 61 negative. For male Ss, 47 were positive and 46 negative. The investigator wanted to draw conclusions regarding the largest positive values and was quite disappointed when the action of chance was indicated.

A crisp example of the misleading nature of certain relationships reflected in correlations is the figure of .97 reported by Locke (1961) between the number of letters in 29 Ss' last names and certain adjectives descriptive of "personality". It was based on a thorough item analysis and item selection leaving the door open, of course, as Locke planned to backsliding. Ss with longer last names were gay and impulsive, talented and God-fearing, did not smoke or used filters, have more dental fillings, like vodka and have hair of a different color from their fathers'. The reliability of the final list of adjectives, incidentally, was only .67. After maximizing the possibility for regression, Locke found, on cross-validation with an N of 30, an overall correlation of -.80. The initial findings were obviously attributable to maximizing the role of chance.

Then there is the matter of extrapolation. Many popular writers have paid lip service, with due cause, to the dangers inherent in statistical analysis. A book has even been published, entitled "How to Lie with Statistics". One article on this matter had the following section headings: the unspecified average, the biased sample, the improbably precise figure, correlations, the gee-whiz graphs, and semantic tricks.

These are all gimmicks and correct as far as they go. Numbers are slippery things. One has to study them, not take some one else's word for what they add up to. One does not believe graphs that show the 1500 meter Olympic run will clock no time at all in the year 2250, the American ski

jumping distance record to be one mile in the year 2153, not (possibly more plausibly) the Indianapolis Speedway record being 1000 m.p.h. in the year 2397. Another case in point are the height/waistline ratios of Miss America winners over the past 40 years. Weintraub and Eisenberg (1966) have pointed out regarding extrapolation of these figures: "It is obvious that the height/waistline ratio cannot be a linear function of time; women were not wider than they were tall several hundred years ago" (p. 247). Plausibility is one thing; gullibility another.

THE SMALLER THE N THE BETTER

Standard textbooks on statistics (some incorrectly titled "Experimental Design") pontificate the case for massive sampling. Their argument seems to be that the effects of chance are somehow diluted or erased by the magic of masses of information. In the first place, faulty experimental design in the way of failure to control a variable simply multiplies itself with increasing N. Secondly, the more importantly, why should chance operate to a greater extent with small Ns? Chance is not a God peering over E's shoulder saying "I'll make this case deviant, that one average." In a lottery the laws of chance are indifferent to the name of the winner. Chance doesn't work this way. Furthermore, the overwhelming point is that statistical results "significant" on a small number of cases add up to a lot more behaviorally than the same finding with a large sample. A probability of .05 derived from two samples of three cases each means non-overlapping behavioral measures. The same probability accruing to two samples of 300 cases each means little except grossly overlapping distri-

butions with no possibility of individual prediction of behavior. These points do not deny that a large sample fills in the picture of the Universe to a greater extent, but after all the primary focus is on behavioral not statistical principles.

An equally powerful case can be made on the other side of the coin for large samples, large samples, that is, of the behavior of individual organisms. A large sample of behavior from a small sample of Ss coupled with a big impact of the experimental treatment is the American psychologist's dream. Covering a wide range of values of the experimental treatment along with careful selection of a behavioral measure sensitive to the treatment, repeated measurement and replication will head the investigation in the right direction.

THE CASE OF DEVIANT CASES

By dint of studying individual behavior we must be concerned with those cases that fall outside acceptable limits. From a statistical standpoint these stragglers or outliers are a problem; there are dozens of statistical gimmicks and procedures for excluding them from the final analysis. None of them are behaviorally satisfactory however. The investigator, after throwing out such a case, is always left with the gnawing doubt that he has overlooked some angle or other. The problem is particularly pressing when N is small, say three or four cases. The present viewpoint is that these deviant cases may be more important than the nondeviant ones. What stimulus circumstances produced the unusual behavior? The matter hinges in part on the definition of the word "deviant". Here

Here it is taken to mean unusual, infrequent and rare rather than the abnormal implied by it's common usage. Being elected President of the U.S. is infrequent, but would hardly be considered a piece of abnormal behavior in the clinical sense.

The present view is that these unusual cases, particularly in small N studies, should be subjected to careful experimental scrutiny for their own sake. In them may lie the answers to a number of pressing experimental problems. In a similar vein one might wish to investigate the background and current status of the greatest acrobat or pianist in the world. They are certainly deviant in a frequency sense; they occur most rarely.

NOSE COUNTING AND THE BINOMIAL EXPANSION

Probably the simplest and most efficient analytical tool available is the binomial expansion. It can be used any time the design calls for a chance baseline, but usually is used in the 50-50 case. For instance, if we simply wish to know whether learning occurs under a given set of operations, all we need do is count the number of Ss that show the increase in response strength classed as "learning". If five of five Ss respond more frequently after we've applied an experimental treatment, the odds are 1 in 32 against "pure" chance generating our event. The binomial is comprehensible to the layman and has been easily taught to eight year olds.

The binomial is simple and obvious. Anyone can understand it. The rub comes in knowing when to apply it. I have seen a number of instances where investigators have the perfect set-up for this kind of count statistic

and proceed to fall away into complex analyses that tell them far less about what has happened than the binomial would - and take many times as long to apply. Investigators should be primed, and even design their experiments, so that such simple analytical tools as the binomial can be applied in just one small extension beyond inspecting the behavioral outcome of the investigation.

The case so far has been kept to its simplest form. Where there are reversals, e.g., one event in the seven goes in the opposite direction from the other six, simple tables are available in several standard textbooks and detailed tables for small Ns are reproduced in the original Quick and Dirty manuscript.

For illustrative purposes there are reproduced in Table 1 some real-life data deriving from an investigation of skid-row alcoholics. The numbers represent University of Tennessee Deprivation Scale scores which reflect the presence or absence of environmental support from family, friends, job, etc. Individuals scoring high on a drinking scale (see Pascal and Jenkins, 1961) were matched on age, sex, vocation and education with individuals scoring low on this scale. They were then compared on environmental deprivation.

Without any analysis, it is eye-catching and immediately obvious that each alcoholic score is considerable higher than that of his control partner on the Deprivation Scale. As a matter of fact the two distributions do not overlap. Thus 10 in 10 events go in the same direction and the odds of a chance finding are 1/1024 or P of about .001. Nothing could be simpler and no further analysis is needed. It should be noted that [illegible] ana-

Table 1

Alcoholic and control Ss matched by pairs on age, sex, vocation and education, compared on University of Tennessee Deprivation Scale scores.

(Pascal and Jenkins, 1960)

PAIR	ALCOHOLIC	CONTROL
1	10	5
2	12	6
3	12	2
4	10	2
5	14	2
6	13	4
7	12	2
8	14	3
9	12	4
10	8	4
Mean	11.7	3.4

P = 1/1024
.001

lysis and all of its kind focuses on consistency without regard to magnitude. In this instance, consistency is the main point and magnitude is of little consequence. One might devote considerable effort on applying a match-pair t-test to these data, but the outcome would remain the same. This is not saying the effect isn't large; mean differences are of the order of three to one and the two distributions do not overlap a rare "whopper" finding.

Presented in Table 2 are some numbers from an experiment by Carter and Schooler (1949) dealing with "Value, need and other factors in perception". Without belaboring the questionable behavioral status of these terms, the conclusion is drawn that "the rich and poor children's judgments were essentially the same....". This conclusion is incorrect. There are five events (coins) and in every instance the average judgment of the poor children was larger than that of the rich. Five events in the same direction occur only 1/32 times on a chance basis. Thus the consistency looks potentially real although the magnitude is admittedly small. Both sides of the analysis coin - magnitude and consistency - must be examined if the data are to be squeezed dry. In this case, essentially "no difference" was concluded where perfect consistency exists. Instances of this point could be multiplied, but the matter should be clear.

Another case in point involves some data based on Sheldon's somatotype measures and anthropometric variables as they relate to the criterion of success or failure in flight training during World War II. The biserial correlations between his 12 measures and the criterion were as follows:

Table 2

Average judgments of coin size in millimeters by rich and poor children.

(Carter and Schooler, 1949)

	DIME	PENNY	NICKEL	QUARTER	HALFDOLLAR
SIZE	17.8	19.0	21.2	24.1	30.5
Rich	16.3	17.6	21.0	25.4	33.1
Poor	16.5	18.6	21.2	25.7	33.9
t	.5	2.2	.3	.5	.8

-.10	.08
.05	-.03
.11	.06
.03	-.01
-.07	.08
.02	.11

The absolutely low level of these correlations is not surprising since these were highly selected individuals and the distributions were compressed and truncated. The eye-catcher is the pivoting of the numbers around zero. Four are negative and eight positive with a mean of about .027. It seems unlikely that prolonged statistical manipulation will yield much beyond the conclusion of a near chance finding.

Table 3 contains some numbers based on quite complex procedures (Pascal et al, 1966). They represent average ratings over a number of behavioral variables from S's report of the behaviors exhibited by his parents toward him in the early years of his life. In other words, they are a large sample of behavior from a small N. All Ss had surgical intervention for their ulcer symptoms so they are very homogeneous in this regard. Despite this similarity, considerable difference emerged between those who lost their ulcer symptoms after surgery and these who did not. Since matched pairs were involved the binomial analysis can be applied. These numbers were selected because they present complications. In the first instance there is a tie for the average ratings for Pair 7 for the stimulus category "Mother". By reference to the appropriate binomial table the chances are 11/1024 of getting nine events in 10 in the same direction

Table 3

Pascal-Jenkins Scale ratings for Mother and Father for 10 pairs of Ulcer patients matched on sex, age, vocation and education, one member responding successfully to ulcer surgery, the other failing.

(Pascal et al, 1966)

PAIR	MOTHER		FATHER	
	Success	Failure	Success	Failure
1	2.8	1.5	1.0	1.6
2	2.8	1.9	2.7	1.8
3	2.7	2.4	2.8	2.3
4	3.0	1.9	3.0	1.6
5	2.9	2.7	2.2	2.2
6	2.6	2.6	2.0	2.2
7	2.9	2.9	2.9	2.7
8	2.8	2.1	2.3	1.4
9	2.6	2.1	2.2	2.3
10	3.0	1.8	3.0	2.0
Mean	2.9	2.2	2.4	2.0

The chances of 10/10 are 1/1024. The tie is split in half by averaging these two probabilities with the outcome's being 6/1024. One could, of course, throw all ties against oneself, but this seems like too much deck stacking.

The "Father" case for Pair 7 is even more complicated, there being three reversals and one tie. The tie is treated as in the previous case. The chances of obtaining 10/10 events in the same direction are 1/1024, 9 are 10/1024, 8 are 45/1024, 7 are 120/1024 and 6 are 210/1024. Remembering that we always want the probability of an event as extreme or more extreme, the total probability for 7 or more events is 176/1024 while the odds for 6 or more events are 386/1024. Averaging out these last two figures (176 and 386) we obtain an overall figure of 281/1024. About 280 times in a 1000 chance would produce a result like this.

What all this verbiage and artful number management adds up to is what one can see with the naked eye: there is really only a slight difference between Successes and Failures as regards "Father". (In defence of the investigation, these are the "worst" set of data selected from a number of experiments.) Again, the point is clear. Without any particular statistical sophistication, one can scan a complex set of data and see what's happened to the point of drawing the appropriate and relevant conclusion. Undoubtedly it takes practice. Reasonable advice calls for looking at the numbers of published papers, not the words.

NOSE COUNTING, ASSOCIATION AND CORRELATION

The case of the beer tax facts. One of the items that triggered off

this return trip to quick and dirty statistics never-never land is presented as Table 4. Quick inspection of it, particularly the first and last numerical columns, suggests a substantial relationship of a negative nature: the more beer consummed the less the tax, and conversely the higher the tax, the less the beer drunk. (The pamphlet accompanying the table argues the unfairness of the case, but we are not concerned here with economics.) There are over 450 numbers in this table. That is too many to analyze unless one is practicing arithmetic. The case is an excellent one for applying and demonstrating short-cut procedures.

Suppose we're simply interested in determining whether this apparent negative relationship between taxes and beer consumption is "real", i.e., is large enough to provide a base for arguing a change in taxation. Further, suppose we're interested in the overall tax structure and not local matters, and finally suppose we're not good at arithmetic and thus want to work with as few numbers as possible to minimize the possibility of error. The solution is simple: take the extreme cases from the first and last column. If the relationship holds in this sub-sample of data, it should hold across the board.

The only gimmick to watch for here is a curvilinear, say U-shaped, relationship where we happen to select data that fits a straight line portion of the relationship. (Graphical representation obviously helps in this regard.) Inspection clearly indicated no changes in direction in trend in the numbers presented. One should always remember that correlation is nothing more than a number reflecting to what extent high numbers in one set go with high (or low) numbers in the other set. The way to find out is to look and see.

Table 4

1964 BEER TAX COMPARISON CHART BY STATES

States Listed According to Total Tax Rate		State, City, County Total Tax		State Tax		City and County Tax		Retail Sales Tax		Per Cap Consump By Gallon
		Barrel	Case	Barrel	Case	Barrel	Case	State Beer Sales Tax	Local Beer Sales Tax	
1	So. Carolina	$19.84	$1.44	$19.84	$1.44	0	0	3%	0	7.1
2	Georgia **	14.88+	1.08+	14.88	1.08	No Limit	No Limit	3%	0	7.1
3	Alabama **	13.23+	.96+	13.23	.96	No Limit	No Limit	4%	1%	5.9
4	Mississippi	13.23	.96	13.23	.96	0	0	3½%	1%	6.3
5	Tennessee **	11.80	.86	3.40	.25	$8.40	.61	3%	1%	9.2
6	No. Carolina	11.57	.84	11.57	.84	0	0	3%	0	6.8
7	Florida	11.57	.84	11.57	.84	0	0	3%	0	14.3
8	Louisiana **	10.00 to 11.50	.73 to .84	10.00	.73	to $1.50	to .11	2%	1%	14.6
9	Oklahoma	10.00	.73	10.00	.73	0	0	2%	[illegible]	8.9
10	Virginia	8.27	.60	8.27	.60	0	0	0	0	12.7
11	So. Dakota *	4.00 & 8.00	.29 & .58	4.00 & 8.00	.29 & .58	0	0	2%	0	11.4
12	Alaska	7.75	.56	7.75	.56	0	0	0	3%	13.3
13	Maine	7.75	.56	7.75	.56	0	0	4%	0	15.2
14	Michigan	6.61	.48	6.61	.48	0	0	4%	0	20.9
15	Vermont	6.20	.45	6.20	.45	0	0	0	0	16.7
16	W. Virginia	5.51	.40	5.51	.40	0	0	3%	0	11.0
17	Texas *	4.30 & 5.12	.31 & .37	4.30 & 5.12	.31 & .37	0	0	2%	0	16.2
18	Arkansas	5.00	.35	5.00	.35	0	0	3%	0	7.6
19	Ohio	4.96	.36	4.96	.36	0	0	3%	0	18.1
20	No. Dakota	4.96	.36	4.96	.36	0	0	2½%	0	15.9
21	Idaho	4.65	.34	4.65	.34	0	0	3%	0	13.8
22	Utah *	1.10 & 4.00	.08 & .29	1.10 & 4.00	.08 & .29	0	0	3%	1%	8.8
23	Kansas *	3.11 & 3.72	.23 & .27	3.11 & 3.72	.23 & .27	0	0	3%	0	10.1
24	New Hampshire	3.72	.27	3.72	.27	0	0	0	0	21.1
25	Pennsylvania	3.31	.24	3.31	.24	0	0	5%	0	19.1
26	Minesota *	1.60 & 3.20	.12 & .23	1.60 & 3.20	.12 & .23	0	0	0	0	16.7
27	Indiana	2.71	.20	2.71	.20	0	0	2%	0	15.3
28	Kentucky	2.50	.18	2.50	.18	0	0	3%	0	12.3
29	Arizona	2.48	.18	2.48	.18	0	0	3%	1%	16.7
30	Iowa	2.48	.18	2.48	.18	0	0	2%	0	15.6
31	New Mexico	2.48	.18	2.48	.18	0	0	3%	1%	13.9
32	Nebraska	2.48	.18	2.48	.18	0	0	0	0	17.5
33	Connecticut	2.07	.15	2.07	.15	0	0	3½%	0	15.6
34	Delaware	2.00	.15	2.00	.15	0	0	0	0	17.1
35	Massachusetts	2.00	.15	2.00	.15	0	0	0	0	16.8
36	Colorado	1.86	.14	1.86	.14	0	0	3%	2%	14.6
37	Illinois	1.86	.14	1.86	.14	[illegible]	0	3½%	½%	19.3
38	Nevada	1.86	.14	1.86	.14	0	0	2%	0	25.0
39	Rhode Island	1.55	.11	1.55	.11	0	0	3½%	0	19.8
40	Dist. of Columbia	1.50	.11	1.50	.11	0	0	0	3%	22.5
41	Montana	1.50	.11	1.50	.11	0	0	0	0	19.7
42	Washington	1.50	.11	1.50	.11	0	0	4.2%	0	15.2
43	New York	1.38	.10	1.38	.10	0	0	2%	3%	19.0
44	Oregon	1.30	.09	1.30	.09	0	0	0	0	15.6
45	California	1.24	.09	1.24	.09	0	0	3%	1%	15.6
46	New Jersey	1.03	.08	1.03	.08	0	0	0	3%	18.9
47	Wisconsin	1.00	.07	1.00	.07	0	0	3%	0	26.6
48	Maryland	.93	.07	.93	.07	0	0	3%	0	19.6
49	Missouri	.93	.07	.93	.07	0	0	3%	0	17.1
50	Wyoming	.62	.05	.62	.05	0	0	2½%	0	14.8
51	Hawaii ***	20% of wholesale price		20% of wholesale price		0	0	4%	0	10.9

Table 5 presents comsumption figures for the nine states with the highest and the nine with lowest tax rates. (Nine is obviously arbitrary; it's small enought to simplify arithmetic and the middle score of an odd number of measures constitutes an average.) A brief examination of the table reveals non-overlapping distributions: the highest number in the first column is less than the lowest number in the second column. For those interested in a slightly more sophisticated treatment, a permutation-combination analysis of 10 events beating 10 others yields a probability around five in a million, a quite rare occurence on any basis. (Behaviorally speaking, the binomial expansion bears on matched or paired events so it is not applicable to these two sets of 10 independent events.) In any case, these data reinforce the point that a great deal can be read into results by careful inspection.

Numbers are sometimes useful in summarizing findings. In this instance it would be handy to have a single number to represent the intensity of relationship, association or correlation between the two dimensions of variation, tax rate and beer consumption. The easiest way to obtain such a number is to sort the data into a two-by-two table. Table 6 represents this transformation for the data of Table 5. The grand mean (mean of means) was taken for the two columns of Table 5 and the individual cases sorted as above or below this value while retaining the original classification of high or low tax. A Phi coefficient has been computed for the resulting two-by-two sort in Table 6. Phi is easy to compute and represents the more elaborate correlation coefficients quite accurately. It consists of a fraction, the numerator of which is the difference between

Table 5

BEER TAXES AND CONSUMPTION

Per capita consumption in gallons by the nine states with the highest and lowest total tax per barrel.

	CONSUMPTION	
	Highest Tax	Lowest Tax
	7.1	15.2
	7.1	19.0
	5.9	15.6
	6.3	15.6
	9.2	18.9
	6.8	26.6
	14.3	19.6
	14.6	17.1
	8.9	14.8
Median	7.1	17.1
Mean	8.9	18.0
Grand Mean	13.5	

Table 6

COMPUTATION OF A CORRELATION COEFFICIENT (PHI) FOR THE BEER TAX DATA OF TABLE 5

	GREATER THAN GRAND MEAN OF 13.5	LESS THAN GRAND MEAN OF 13.5	TOTAL
High Tax	2	7	9
Low Tax	9	0	9
Total	11	7	18

$$PHI = \frac{(2)(0) - (9)(7)}{\sqrt{(9)(9)\quad(11)(7)}} = \frac{-63}{79} = -.80$$

the products of the diagonal numbers. The denominator is the square root of the products of the four marginal totals. In this instance a figure of -.80 emerges indicating a substantial negative relationship between tax rate and beer consumption and, more importantly, clearly supporting the inspectional conclusion.

The usual word of caution is called for regarding the interpretation of indices of correlation and association. The easy part is computation; the hard part is saying what the produce means. Things go together or covary. One does not "cause" the other. There may be a substantial correlation between the abortion rate in Brooklyn and the rainfall in Rangoon, but it would be difficult to uncover a cause-effect relationship. In other words, caveat emptor when it comes to the interpretation of correlations and other measures of association. For example, Sargent (1955) computed the correlation between the number of letters in the names of the months and the mean monthly precipitation for 1947. The figure was -.61 with an associated probability of less than .05. The reader is left to figure out what the covariation means.

Thus far we have dealt with instances of nice, clearcut positive findings. Inspectional analysis applies equally effectively to negative results or cases of "essentially no difference". A case in point comes to hand in the way of a study of activity patterns of schizophrenic patients (Chapple et al, 1963). Among many other things, the investigators were attempting to be behaviorally economical in seeing if four observations per day would suffice instead of six. The differences are presented Table 7 S by S, for four separate days.

Table 7

Differences in activity between six and four observations per day for 10 schizophrenic patients.

(Chapple, 1963)

S	DIFFERENCES	S	DIFFERENCES
1	1	6	-9
	5		4
	-1		2
	-2		11
2	0	7	0
	5		7
	0		-2
	0		-6
3	-8	8	1
	1		-2
	9		18
	-2		4
4	6	9	0
	-2		0
	-4		-5
	-1		-2
5	-4	10	0
	6		-1
	-13		0
	-7		0

There are many ways of cutting and slicing these data. The investigators did it the hard way by doing 10 individual t-tests, one for each S. Simple counting of pluses, minuses and zeroes reveals a 14, 17 and 9 split for the forty numbers, a finding quite in accord with chance expectation. Inspection of the data suggests no large systematic differences; counting supports this view. If one wishes to be a little more thorough about the analysis, a total can be taken, S by S. Five Ss show negative sums, four positive and one zero. The mean difference is 0.9. Again chance prevails by this token.

At this juncture it seems wise to comment that there are in the behavioral world some sets of data that are too complex to be handled by inspectional analysis. Factor analytic studies are a case in point. Data in behavioral science seem to be more complex the less we know. As knowledge increases, simplicity sets in and the stage is set for once-over-lightly kinds of analysis such as inspection. In any event, there is a serious question concerning the utility of factor analysis and similar cumbersome procedures. They may be a defense, an escape through the machine for the investigator, but they help the audience little. I believe that at least one expert in the field said that no worthwhile test has ever been developed as a result of a factor analytic study. This sounds reasonable.

Returning to the main stream of this section, some data are shown in Table 8 having to do with Experimenter differences. Four different E's each tested four pre-school children in a discrimination learning set-up. If the child had not learned in 36 trials, testing was ended. It is

Table 8

Trials to reach a criterion in discrimination learning for four sets of four pre-school children each tested by a different Experimenter. Learning was terminated after 36 trials.

	EXPERIMENTER			
S	1	2	3	4
1	8	20	36	36
2	6	17	5	36
3	12	5	32	28
4	23	12	15	10

clear that E - 3 had one S fail to learn while E - 4 had two. Are the four differences large enough to warrant taking action? While there is clearly a suggestion that E - 4 has more difficulty in conditioning children in this situation, no hard conclusion can be drawn. Furthermore, it hardly seems necessary to apply Chi Square or other indices of frequency difference. Inspection makes E - 4 stand out. The real question is whether he continues to stand out with repeated testing. If two more sets of four children each were subjected to the discrimination operations by each E and the same trend emerged, it would be most plausible to consider E - 4 as being drawn from a different universe than the other examiners and to study him as the variable in the differential findings.

Unfortunately, the behavioral literature is replete with positive examples amenable to inspectional analysis, but the bias about publishing negative findings on the part of both the author and the editor cuts way back the instances of negative findings or small, inconsistent, insignificant results. Negative findings are on many occasions more important than positive ones - they allow us, for instance, to disregard variables. A Journal of Negative Findings is still needed.

There are a huge number of tests of association, contingency and correlation - far to many to even mention in this context. If one wishes to use one, the Fisher-Yates Exact Test is recommended for the two-by-two set-up. It corresponds to the Phi Coefficient although it yields only direct probabilities with no direct indication of extent of relationship. It is cumbersome to compute and Chi Square is a fair approximation to it and much easier to calculate. Across the board, the Phi Coefficient does the job.

HOW BIG? MAGNITUDE CONSIDERATIONS

While statistical procedures are continuous and to a large extent independent of experimental design (although determined by it) - the t-test flows into the F-test, one dimension of experimental variation shades over into two and more - it is practically convenient to separate out the operations for two groups from those for more than two groups with one experimental treatment and, in turn, the latter from situations involving two or more dimensions of experimental variation. Such a course will be followed here. In addition, while related, the statistical procedures for two independent groups differ from those for two related groups and will be further separated.

The outline follows of the subsequent sections of this paper dealing with the statistical assessment of magnitude for the several types of experimental design in increasing order of complexity.

I. The two group case: independent groups, matched pairs or self-control design and matched groups.

II. Anova: one dimension of experimental variation involving three or more groups or conditions.

III. Anova: two "simultaneous" dimensions of experimental variation: matching or self-control, repeated measurement and "simple" factorial design.

IV. Complex Anova: more than two "simultaneous" dimensions of experimental variation.

The emphasis in discussing these procedures, consistent with the rest of the paper, will be on easy, error-minimizing, efficient, short-cut

ways of treating data and clarifying trends clearly visible in behavioral measurements.

MAGNITUDE: THE CASE OF TWO GROUPS

1. Independent Groups. The overlap in the various statistical approaches is indicated by the fact that several sets of data appropriate to this section have been presented in other contexts. For purposes of exposition, Table 9 is given in which some modified findings are summarized.

There are a good half dozen ways to tackle these numbers statistically, but, as always, inspectional analysis is numero uno. Significance of some kind is clearly shown by the fact of overlap of the two sets of five numbers by only one case. The t-test would appear to be the most appropriate analytical technique, but it is the most insensitive, yielding a P of only .055 while the Arrangement Technique (diluting the difference by putting the tied case for high SES first) produces a value of .008. Sorting the data above and below the grand mean (31.5) yields a Fisher-Yates P of .024 with a corresponding Phi Coefficient of about .82. (Considering time, it took about one minute each for the Fisher-Yates and Phi and nearly five nimutes for t-test.) By any token the experimental treatment of SES has had a large impact on ability to reverse in discrimination formation.

In the previous Q & D paper, considerable space was devoted to the Range Test. It is one of many variants of the t- and F-tests based on substituting the range for the standard deviation. The usual caution applies: beward of extreme outliers; a single deviant case can produce in-

Table 9

Trials to a criterion in discrimination reversal as a function of socio-economic status in pre-school children (hypothetical, doctored data based on preliminary findings).

S	LOW SOCIO-ECONOMIC STATUS	HIGH SOCIO-ECONOMIC STATUS
1	40	32
2	38	30
3	36	28
4	34	25
5	32	20
Mean	36	27

Phi = .816

P for Arrangement Technique = .008

P for t-test = .055

P for Fisher-Yates Exact Test = .024

significance where significance really exists. The Range Test is simple, efficient and easily understood. It consists of taking the range across means (for two or more values), multiplying it by the (average) number of cases in the samples and dividing the resulting figure by the average range in the sample data. In Table 9, the range across means is 9, N is 5 and the mean range in the samples is 10. The ensuing Range Test value is 4.5. For two groups and Ns of 10 or less, the resulting value can be referred to the t-table. For more than two conditions and Ns larger than 10, degrees of freedom are computed by multiplying the average number of cases minus two by the number of conditions and referring the resulting Range Test figure to a special table contained in the previous Q & D manuscript.

In any event, it is obvious that the Range Test is far more significant and far less time-consuming than the conventional t-test. The P-value for the data of Table 9 by this technique is about .001, contrasted to the .005 according to the classical t-test.

The range is a highly useful estimate of variability so long as grossly deviant cases are not involved. For example, the range divided by N is a close estimate of the standard error of the mean when outliers are not involved and short cuts a good deal of computational labor in deriving the t-test value.

Across the board, the data must be carefully examined and the one or two most efficient procedures applied, i.e., those that maximize return from the behavioral data and minimize labor and error.

2. Matched Pair or Self-Control Design. This variant of the two

group case is far and away the most efficient if the matching measures correlate with the experimental behavior so that variability is cut down. A case in point is shown in Table 10 where doctored data make the point. Examination of the data shows the E - group exceeding the C - group by a small margin. Treated as independent groups, the P-value emerging from the application of the t-test is .18. It is, however, obvious to inspection that a substantial relationship holds between the two sets of numbers. As a matter of fact, the rank correlation is .88. Another obvious point is that six of the eight differences are positive. The Binomial Expansion, previously treated in detail, is clearly applicable to these data, but yields a P-value of only .144. It is to be noted that the two reversals are the smallest in absolute magnitude. This situation calls for a test sensitive to these magnitudes. The Wilcoxon Rant T-Test is appropriate. It involved ranking the differences by magnitude without regard to sign and sorting out sums of ranks by signs. The smaller sum of ranks is then referred to the table presented in the previous Q & D paper and in some standard statistics texts. The resulting P is ca. .02. This is probably as dry as the data can be squeezed, but to complete the picture, classical t was applied and produced a P of .018.

There are several points here. The first is to match on variables that have something to do with behavior in the experimental situation so that a correlation in performance is generated. If little relationship is produced, time has been wasted in the matching procedure.

The self-control design is, of course, the limiting and best case of matching since each S is more like himself on different occasions than

Table 10

Hypothetical data: The efficiency of matching or self-controlling versus independent groups.

S	E	C	DIFF
1	20	16	4
2	34	35	-1
3	24	22	2
4	37	29	8
5	23	24	-1
6	35	30	5
7	30	27	3
8	29	25	4
Mean	29	26	3

he is like anyone else. Another point is that if you've matched and a correlation has come about to cut down on variability, by all means take advantage of it by applying the statistical procedures appropriate to the set-up. It is clear in Table 10 that a correlation has emerged as reflected in the greatly decreased variability in the distribution of difference scores as contrasted to the spread in the original measures. Thus a matched-pair treatment is called for, the binomial for consistency and rank T and/or classical t for magnitude. Whenever the reversals are small in size, the latter techniques - that take magnitude into account - are preferrable to the straight count procedure.

Another situation where magnitude treatment is needed involves very small Ns. For instance, in an experiment on the combined application of reward and punishment in conditioning on extinction responding, two pairs of pigeons, operating in standard Skinner boxes, were matched on APR responding prior to the use of electric shock. One member of each pair was shocked until responding stabilized at circa 5% of its original value. Then extinction operations were applied. Total extinction responses in 11 hours were:

Pair	Shock	Non-Shock
I	640	13,690
II	370	13,160

One hardly need analyze these data; they serve as a tour de force. The classical t-test is the only analytical procedure applicable and it yields a P-value of .004 for the one degree of freedom involved - if one is a stickler for statistical protocol. Actually, no analysis is neces-

sary and each pair of birds should be considered a separate experiment. The point is made.

Another, somewhat more dramatic example of the same point is contained in an experiment in crowding the threshold in ECT. Patients exposed to Electro-Convulsive Treatment exhibit some resistance to the procedure, a small part of which shows up in delay in insertion of the tongue depressor that is used to prevent tongue swallowing during convulsions. A student of mine was interested in crowding the threshold on this delay. He first took "before" measurements, a kind of latency of depressor insertion. This interval in sec. for the experimental Ss to be trained was 12, 30 and 11. They were paired with controls with intervals of 5, 14 and 10. (The cards were deliberately stacked against the treatment by having shorter latencies for Control Ss.) The experimental treatment consisted of putting dissimilar objects in the mouths of Experimental Ss and gradually, keeping the latency short, increasing similarity to the tongue depressor. Then tests were conducted in the ECT setting. The "after" scores in sec. for the Experimental Ss were 1, 9 and 2; for the Controls 13, 30 and 19.

Before turning to the actual treatment of the numbers, let's look at the overall design picture. This experiment can be looked on in a quite complex way - over and beyond the complicated context in which it is set. One could argue, admittedly, somewhat irrationally, for an analysis of covariance in which the pre-treatment measurements were partialled out of the post-treatment ones by considering the covariation between pre and post treatment scores. Setting aside the question of whether correlation based on three points means anything, the question remains, is com-

plex analysis worth the trouble and will it yield anything beyond what is produced by simple analysis? The answer, of course, is no. As a matter of fact no consideration was given to standard analysis of covariance, but rather the straight forward procedure was followed of converting the latency scores into percentage change scores or savings scores from "before" to "after". These turned out to be:

PAIR	EXP.	CONTR.
1	92%	-160%
2	70%	-114%
3	82%	-90%

The pairing now becomes almost irrelevant because of the size of the effect. We have two sets of three events failing by a large margin to overlap with three other events and P is .05 by the Arrangement Technique. For didactic purposes the t-test was applied to the distribution of three differences across pairs and yielded a P-value of .008. In this case, the training had such a large impact that the correlational feature built in by matching was washed out. As a matter of fact the P based on an independent sample t-test is slightly smaller than that occurring to the matched t. In passing it might be noted that the Range Test is not appropriate to these data because of the great disparity in the sample ranges, i.e., 22% versus 70%, but the outcome is consistent with the findings from the other procedures.

Again, the reader is advised that the purpose of statistics is to "prove something" - the something his naked eye tells him has occurred in

the behavior of his organisms. The "telling" is, of course a matter of discrimination and, like all discrimination formation, takes time and practice. Remember not to bother to analyze if "nothing" has happened, if little or no behavioral differential between the groups is clearly apparent.

The data of several of the tables presented earlier in this report are amenable to examination by the techniques spelled out in this section. It might be worthwhile to look at those numbers in this light.

3. Matched Groups. On occasion it is possible to reap experimental and statistical benefits from group matching where individual pairing is not possible. In group matching, equivalence is achieved in the mean and standard deviation of some a priori measure known or thought to correlate with behavior in the experimental situation. It is a less precise and sensitive measure than pairing which in turn is less exact than use of the self-control procedure. If, however, behavior on the group matching variable relates to the experimental measurement, there is a cut back in variability and a corresponding gain in statistical sensitivity and precision, i.e., the P-value is decreased. Group matching is employed for several reasons. Among them are large Ns where pairing is overly time-consuming; time limitations where Ss, say, go directly from conditioning into extinction and time does not permit matching and loss for some reason of one number of an already matched pair.

The statisitcal procedures for analyzing data by the matched group technique are spelled out in most statistical textbooks. Here, suffice it to say that the overall correlation for both E and C groups combined

is computed between the "before" and "after" measurements. There is one real potential gimmick in computing such a correlation. By the nature of the experimental treatment, it sometimes happens that the relationship between the matching measure and the criterion is thrown off by the treatment so that differential correlation across the E and C groups emerge. I have seen data where r is .80 in the C-group and near zero in the E-group. In such cases pooling the numbers for correlational purposes appears questionable. One could argue for computing the correlation separately and combining correlations by z-transformations, but this seems to be a rather sticky refinement. The investigator must decide whether to forgo his matching in cases such as this or simply report the differential correlation and go ahead and combine anyway in order to gain whatever precision and increase in sensitivity accrues to the matching. In any event, he is obligated to examine closely the relationship between the two variables separately for the E and C groups. (This matter will be considered again in connection with the analysis of covariance in a later section.)

An example of the use of the group matching procedure is contained in an experiment dealing with the hors d'oeuvre effect of prefeeding pegeons operating in a Skinner box. Initially, the design called for 12 pairs of pigeons matched on responding in conditioning to be divided into two experiments of six pairs each. One pigeon was ailing and did not complete conditioning and had to be dropped from the experiment. Fortunately, this bird was near the middle of the distribution, so rather than discard his partner, group matching was used.

The experimental treatment consisted of pre-feeding 11 of the 23

birds an amount of food that increased their body weight approximately 1.5% prior to an extinction test. The experiment aimed at one test of the drive-reduction reinforcement position, which assumes that over a wide range, increased drive leads to increased response strength. The contrary, contiguity position adopted in this experiment was the reinstating cues (food) associated with responding in conditioning would increase response strength. Thus, by this token, increasing body weight (decreasing drive) by pre-feeding prior to extinction test would provide more of the stimulus compound associated with responding during previous conditioning and thereby generate more responses in the extinction test. In a crude sense we were trying to "prove" the Null Hypothesis associated with the drive-reduction position, i.e., show no difference. A lack of difference would, of course, favor the contiguity cue-reinstatement view. A difference favoring the lower-drive, prefed group would be gravy. The latter was the outcome as shown in Table 11 were the distribution statistics are presented.

The matching correlation between responding in conditioning and the 10 min. extinction test was .65 with no differential effects appearing across E and C conditions. Such intensity of relationship appreciably reduced the standard error of the difference so that a one-tailed P appeared of .10 favoring the prefed group and the cue reinstatement hypothesis.

That this effect is "real" is demonstrated in the fact that a number of other experiments yielded comparable results with some even more striking. In one, for example, where pairing was achieved, eight prefed birds wxceeded their control partners. In these experimenters such variables were introduced as amoung prefed, time lag between pre-feeding and

Table 11

The hors d' oeuvre effect: The influence of pre-feeding on 10 min. of non-reinforced Skinner box responding in pigeons.

	Hors d' oeuvre		No Hors d' oeuvre
N	11		12
$\bar{X}$	96.0		68.0
SD	69.8		59.2
t		1.3	
P		.10	

test and schedule of reinforcement.

Group matching is not a particularly common practice. Where one can group match, he can usually pair - a far more efficient technique. Furthermore, if behavior on the matching dimension does not correlate substantially with the experimental behavior the procedure is a waste of time. Also, differential relations between E and C must be considered. Sometimes matching is too much trouble, particularly where N is huge. On a few occasions, as the one cited, it's worthwhile.

MAGNITUDE: THE CASE OF THREE OR MORE GROUPS WITH ONE DIMENSION OF EXPERIMENTAL VARIATION: SINGLE CLASSIFICATION ANOVA

The continuity between this situation and the case of two independent groups, between the t- and F-tests, has already been indicated. Standard textbooks spell it out; it need not be stressed here. In many instances we are experimentally interested in a functional relationship between degrees of treatment and behavior. Thus we employ three or more points of our experimental variation and corresponding groups. This presents a situation appropriate to one-dimensional or single classification analysis of variance. The complexity of anova lies in the increased N and nothing else. Basically, it's nothing but an elaborated t-test involving a comparison of treatment differences across conditions with an overall estimate of S-to-S variability ("individual differences"), that is, a ratio of variation in means to variations across individuals. As will be indicated, there are easier ways than the traditional for accomplishing this.

Before launching into a treatment of single-classification anova, a basic word of caution is needed. When significance is achieved the procedure does not indicate what aspects of the behavior or what conditions generated the significance. In other words, the outcome of the application of anova to data is an open-ended proposition. For a given level of significance of F, the functional relationship can be linear, exponential or parabolic. Anova doesn't "care". Additional tests of significance (as well as careful scrutiny as always) are called for to tease out the exact features of the data producing the significance. Fortunately, tests are available for detecting outlying means that help to pin down the significance, but it should be indicated that they are cumbersome from an arithmetic standpoint. More will be siad on this point later.

1. Rank anova. The best way to illustrate anova is by an example. Some actual data are presented in Table 12 that concern the gross bodily activity of rats in an open field at three different drive levels determined by percent of satiated body weight. Gross movements were defined in terms of eight-inch square traversed and rearing responses. The overall project dealt with the impact of novel, unfamiliar stimuli of varying intensities and characteristics on performance of gross and fine movements.

The first item to be spotted (after noting the clear trend for gross movement to increase with drive) is the outlying case in the 90% group which tops all others in responses. The ensuing heterogeneity of variance poses real problems for classical anova and also for the Range Test considered in the previous section. The classical F test can be applied to

Table 12

Number of gross movements (locomotion and rearing) emitted in 5 min. by three groups of rats at different drive level.

	DRIVE LEVEL		
S	80%	90%	100%
1	154	114	108
2	172	217	127
3	204	87	97
4	139	128	127
5	181	145	103
6	165	—	178
7	138	—	—
Mean	164.6	138.2	123.3
Median	165.0	128.0	117.5
Range	66	130	81

the variances as the ratio of the larger to the smaller variance, but more appropriately the Hartley F-maximum test should be used. (It is treated in most standard statistics textbooks.) While its value only reaches the 10% level, problems remain for anova procedures that deal with the raw hetereogeneous numbers.

A simple way out is to transform the raw scores to ranks and apply the Kruskal-Wallis Rank Anova as spelled out in most current standard statistics texts. Note that ranking the data tends to minimize heterogeneity of the numbers, it does not change their relative standing. This technique has the disadvantage along with the traditional anova, of allowing opportunity for considerable arithmetical error, but it is still the most appropriate procedure for the numbers at hand. The essence of this procedure is to pool all the numbers and rank them from, say, high to low, sum ranks by columns and substitute the sums of ranks into a formula which produces a number, treated as a Chi Square, that reveals whether the column sums have pulled sufficiently apart to warrant rejection of the Null hypothesis of a common target or parent population. In this instance the overall P-value from the rank anova is .027.

None of the anova procedures pinpoints what features of the data are generating the significance. In the current instance, inspection suggests the 80% group to be deviant with the behavior of the other two groups tailing off in a curvilinear, asymptotic fashion. The data are probably too crude to warrant more refined statistical treatment. The point is clearly made that higher drive tends to be associated with greater gross bodily movement.

Table 13

Rats' Skinner box extinction responses with 24 hr. food deprivation in extinction and the given hours of deprivation at conditioning.

(Finan, 1940)

	HOURS OF DEPRIVATION IN CONDITIONING			
	1	12	24	48
N	28	29	30	30
Mean	31.6	62.0	53.8	45.6
Median	25.0	57.5	40.0	41.0
SD	25.4	35.0	41.8	18.2
Estimated Range	100	170	190	60

2. Classical Anova. To illustrate the use of classical anova and the Range Test, some data obtained by Finan (1940) are presented in Table 13. Before turning to the analysis, let's consider this experiment from a design and behavioral standpoint. In essence, what Finan did was condition four groups of rats in a Skinner Box at 1, 12, 24 and 48 hours of food deprivation. All were then extinguished at 24 hours of deprivation. This set-up becomes an incomplete block design where the complete design would have all four deprivation values represented in extinction as well as in conditioning. The absence of complete information thus limits the inferences that can be drawn.

Given the set-up as it is, certain a priori considerations apply. The fact of the matter is that drive was changed from conditioning to extinction for three of the four groups and not changed for the forth. The principle of generalization and its correlary of generalization decrement clearly apply: The greater the change in the stimulus conditions, the greater the behavioral decrement. On the face of it the groups with the greatest change in drive should show the greatest response decrement - and they do. The 1 and 48 hour groups are below the level of the 12 and 24. The situation is complicated by some special drive manipulations Finan employed and even more by the fact, shown in the data of Table 12, that higher drive leads to increased bodily activity which, in this instance could readily be channeled into the bar pressing response. The effect is there; the responses of the 48 hour group exceed those of the 1 hour group with both roughly equidistant from the 24 hour group in deprivation. All in all, the generalization position fits the data nicely except for the peak performance of the 12 hour group and this may well be sampling or

attributable to the special operations.

The generalization principle provides a logical and legitimate basis for combining the 1- and 48-hour groups against the pooled 12 and 24 hour groups. Ardent and avid statisticians may throw up their hands and call this a sticky procedure, but behavior theory dictates it. The P-value for the t-test applied to these combined data is .008 suggesting the operation of a systematic variable, namely, generalization and generalization decrement from conditioning to extinction on the drive dimension. In other words, the less the drive change, the higher the level of extinction performance.

Anova is basically a simple though cumbersome procedure. In essence, the deviations or differences across means are compared with chance variation as reflected in differences among individuals. The calculating procedure follows directly: deviations of means around the grand mean are contrasted with the total of individual deviations around means of columns or conditions. The exact calculating steps in deviation or raw score units are treated in all books considering anova and need not be detailed here.

Since Finan presents means and standard deviations by conditions along with a dot graph representing individual performance, the stage is set for the application of classical anova and the Range Test. Following through on the anova steps and disregarding the potential heterogeneity of variance across conditions, yields a P-value of .009 that indicates, by all ordinary standards, enough divergence from chance to warrant rejection of the Null hypothesis. The follow-up analysis by the t-test

supporting the generalization hypothesis concerning these data has already been mentioned. The next step, as always, is follow-up experimentation. A number of such studies (Jenkins, 1955) supports the conclusion that drive change, like any other operational, experimental change, produces response decrement, except for the point already noted that substantial increases in drive lead to increases in gross bodily activity that may be channeled into the recorded response so as to compensate for the change effect.

It's obvious that the anova procedure applied to the two-group as well as the situation involving three or more groups. There are many occasions where it is profitable to pivot experimental findings from one investigation on control data gathered in another experimental setting using the principle of dual controls. In other cases one control group may be the pivot point for several experimental groups. In all instances, by definition, replication is involved. Some pertinent data from an experiment on crowding the threshold with pigeons follow:

S	CONTROL	"INTERNAL" CROWDING"	"INTERNAL-EXTERNAL CROWDING
1	1970	230	580
2	2300	1090	1040
3	3800	2030	1470

In the "Internal" procedure, after conditioning at 80% of satiated body weight, pigeons were completely satiated and then their body weight then very gradually reduced to its original 80% level while exposure to

the Skinner Boxes was continued. In the combined case of "Internal-External" Threshold Crowding, the same procedure was coupled with decreasing the illumination on the pecking window to a minimum and then gradually reinstating the original illumination. The numbers represent extinction responses after the treatment.

First, it is obvious that independent organisms had to be used in the three conditions and second, it is clear that matching could be employed (and was, but will be ignored in this context.) In this apparent anova set-up, the impact of the treatment was large, i.e., crowding the threshold by either procedure cut extinction responding to less than half of that of the control Ss. Only one Experimental S's responding exceeded the lower limit of the Control Ss. One might apply overall anova to these numbers or the t-test to the separate experiments but it's obvious regardless of statistical outcome that behavioral change has occurred. In passing, it might be noted that only the matched t-test is applicable in the pairing case as N is too small for either the Binomial or the Rank T-Test.

3. The Anova Range Test. Since Finan (1940) presented a dot graph indicating individual responses, the range of performance in his four groups can be estimated and is shown in Table 13. The range in the means is a little over 30 responses, N is taken as 29, the mean of the ranges in the samples is about 130 (despite a couple of outlying cases) and the Range ratio value approaches 7.0 with a P-value of less than .01. Here as in the other cases, the Range Test is far easier to apply than the traditional tests and allows for considerably less computational error.

Table 14 presents some data from an auditory deletion experiment

Table 14

Number of items in a message correctly reconstructed by college students as a function of percentage of the message deleted by auditory masking. Maximum correct is 30.

	PERCENT DELETED					
S	10%	20%	40%	50%	60%	70%
1	27	21	18	8	7	7
2	30	22	21	9	13	0
3	28	25	20	8	8	6
4	27	25	13	5	0	2
5	27	25	15	6	11	0
Mean	27.8	23.6	17.4	7.2	7.8	3.0
Mean % Retrieved	92.7%	78.7%	58.0%	24.0%	26.0%	10.0%

that is particularly amenable to the Range Test. In this experiment college students were given instructions to perform with a series of objects placed in front of them. For separate groups, different proportions had been deleted by auditory masking. The score was the number correct of a possible 30 actions. The investigation had to do with the redundancy of the English language.

Examination of Table 14 reveals a large, clearcut trend: the more information deleted, the smaller the number of correct responses. There is a "whopper" effect with the extreme groups differing by a factor of five or more. One might apply classical anova to these results, but it seems like a lot of work when the Range Test will quickly and easily do the job. The range across means is roughly 25 units, N is 5 and the mean range in the samples is about 6.5. The resulting Range value is around 20 with an associated P of considerably less than .01. Extremely high significance is demanded by the inspectional fact that adjacent distributions overlap only slightly except for the 50% and 60% conditions. Inspectional analysis pinned down by graphical representationwould seem sufficient analysis for these clearcut findings.

A comparison of visual and auditory deletion may generalize the case. Whereas in visual deletion of letters in printed material (Jenkins and Mosteller, 1954) with 50% deleted, nearly 90% of the message was correctly reconstructed, here with 50% masked by auditory stimulation less than one-quarter of the message was correctly retrieved. A level comparable to that of visual deletion was found here with only 10% of the message destroyed. The discrepancy is consistent with the view that man is primarily

a visual organism.

Cases could be multiplied ad nauseum illustrating single classification anova, but the point is clear that there are better ways than the traditional ones. The Range procedure is most appropriate so long as one looks for especially outlying cases. Transformation of the data to ranks helps and there seems to be no reason that the range procedure can't be applied to the ranks directly rather than wading through the cumbersome arithmetic of the rank technique. For example, when the gross bodily movement data of Table 12 are transformed to ranks and the Range Test applied to the ranks, a value near the .05 level emerges. In all cases, of course, the more formal analysis should support the trends visible in the data.

After anova, what? Multiple comparisons. As has been noted several times, the outcome of anova can indicate overall significance, but not pinpoint what particular, specific differences are generating this outcome. The essence of demonstrating what a significant anova adds up to lies in teasing apart the means associated with the several conditions. This can clearly be accomplished by inspection of the means and variabilities, but most behavioral scientists require more quantitative evidence. A number of procedures are available (Ryan, 1959), and as such will be noted but not treated. Tukey's Layer Test is one of the better ones where outlying means are peeled off like layers of an onion. The t-test is sometimes used incorrectly. It was developed for testing the hypothesis of zero difference between two and only two means. The distortion introduced when, say, six means are compared and contrasted is apparently large unless a

directional hypothesis was set up on an a priori basis, i.e., mean A predicted greater than B, B greater than C, and so forth. In this instance, however, Mosteller's Testing a Ranking (1950) procedure is far more efficient than other so-called multiple comparisons since, if the means attain the predicted order, the operating hypothesis is accepted without further ado. Nothing could be simpler. The rub comes, however, when the prediction is made and the predicted order of means is not achieved within the limits of sampling variation. Then considerable experimental effort has been wasted. In other words, a large wager is made for a big return, but a loss is also big.

The basic problem in contrasting more than two means in an anova set-up is obtaining an overall estimate of error for any mean that reflects the expected (and obtained) sampling variation in all of the means. Once this parameter is fixed (and one outlying case can create real problems), the procedure is simply one of setting a significance level and determining if adjacent means - arranged in order of magnitude - differ enough to infer separate target populations. The arithmetic is a little lengthy, but the basic notion is straightforward.

Across the board - and this comment applies to forthcoming sections as well as the present one - anova is a handy exploratory instrument where one is not certain what's going on with the numerical patient. It helps one infer overall significance, and, as such, is a systematic operation, but is no substitute for more precise or sensitive analytical tools. It is clearly no replacement for inspection since, as has been already noted, significance can accrue to anova when the relationship is linear, exponential

or parabolic - and behaviorally it usually makes a great deal of difference which it is. Anova, however, doesn't respond to the nature of functional relationships. One other minor objection to anova might be noted. It does not indicate (any more than the t-test or similar measures) the intensity of relationship (far less the direction) involved. Peters and Van Voorhis (1940) (and others since) have proposed a generalized form of curvilinear correlation, Epsilon Squared, as a substitute for anova on the ground that it provides an index of relationship. There is clearly a point here, but the same objections of effort and error apply to this procedure as to anova. There is no substitute for visual scanning and graphical representation as the basic modes of determining the effects of an experimental treatment on behavior.

MAGNITUDE: TWO "SIMULTANEOUS" DIMENSIONS OF EXPERIMENTAL VARIATION, DOUBLE CLASSIFICATION ANOVA

This complicated phrase encompasses three related but disparite situations:

1. Three or more conditions of the experimental treatment with the same Ss rotated through the conditions (self-control procedure) or the use of Ss matched on some a priori basis;

2. The case of "repeated measurements" or "trend analysis" where two independent groups are tested or measured several times over a series of trials or blocks of time;

3. "Simple" factorial design where two experimental treatments are applied "simultaneously" to two or more groups each.

It might be noted, ad initio, that such matters as two "simultaneous" dimensions of variation, be they a correlational element and a treatment or two treatments, are complex matters from the standpoint of both statistics and arithmetic. From a design and experimenting view, they add only slight to moderate additional a priori and experimenting labor and may pay large dividends. It would seem that as design increases a bit in complexity statistics increase geometrically in difficulty. It might be added at this juncture that additional increases in design complexity, such as adding a third experimental treatment, also seems to increase interpretation of the resulting data geometrically. In addition it sets the stage for a major role to be played by one deviant case going against the grain of the group. More will be said on these matters in connection with complex anova. The point to keep in mind is that both statistical and interpretative effort increase greatly as treatments or variables are added.

1. Correlated data. This situation is a variant on the single classification anova theme where the variable added is a correlation across rows by either using the same Ss rotated through the three or more conditions or Ss are matched on a beforehand basis and assigned in trios or larger sets to the several conditions.

As a tour de force in another connection (Jenkins, 1966) I wrote up the following (hypothetical) example of translating everyday business into experimental action.

The Whiff Test. This example stems from the hypnotic state induced by overexposure to TV ads. This attack on the deodorant problem is intended as a rough and ready paradigm for experimental designs dealing with

a comparison of advertised products. The steps spelled out apply equally to detergents, soap, hair tonic, cars, cigarettes, toothpaste, razor blades, dog food, and the like. It might be noted in passing, that these problems are far from trivial in at least one sense: problem significance is met in that the very practical criterion of billions of dollars per year are involved.

The first consideration is, of course, the experimental treatment. This is straightforward. The three deodorants leading in sales are selected for experimental examination. This is an objective and satisfactory criterion for inclusion. Advertising claims as to effectiveness can be ignored since they all amount to the same thing: vague and meaningless come-on. The several deodorants are to applied in equal amounts (or durations) or this property is to be varied systematically as part of the experimental treatment. Also built into the design at this point would be variation in time since bathing and nature of activity preceding application, e.g., social, physical or intellectual.

The core of the design would be to rotate a small sample of Ss, say 10, through all orders of presentation of the deodorants (including a "placebo" and a "nothing" baseline condition) several times, applying a test for odor (The Whiff Test) each time these steps all followed by a replication with 10 more Ss. Subjects should be roughly representative of the target population of deodorant users in age, sex, frequency of use, shaving of axillaries, etc. A sub-sample of non-users might add interesting information.

The dependent variable of behavioral measure is slightly more com-

plicated. While a refined instrument such as Zwaardemaker's Olfactometer could be used to measure odor as a supplement to the proposed test, the latter is simpler and requires no more than the human apparatus. The Whiff Test consists of having three judges without head colds, nasal obstruction or other olfactory difficulties, approach S and sniff (or whiff) at systematic distances from him. Each judge would independently record "yes" or "no" for the presence or absence of odor. Any special features such as intensity or quality of odor would also be noted. Adaptation effects for the judges should be controlled by interpolating periods of nasal inactivity. It is obviously preferable that S not know he is being judged. Information regarding the chemical nature of the deodorants and the amount of perspiration generated by Ss under various conditions is of interest, but not the focal point of the investigation.

Control procedures have already been stipulated for a number of sources of variation. By the self-control design, individual variations are minimized and sensitivity to the treatment maximized. The use of both a "placebo" and a "nothing" condition provides a baseline below which the suppressive effects of the deodorants can be assessed. Mode of presentation, e.g., stick or spray will, of course, be held constant or varied systematically. Other considerations may include training the judges in olfactory discrimination and control of the odor of the deodorants themselves.

Since the culture seems to imbue large numbers of people with reserve - if not fear and anxiety - about numbers and, particularly, about statistical manipulation of them, it seems appropriate to demonstrate the

potentially simple nature of analysis for the types of numbers emerging from this investigation. Numbers, after all are simple and crisp; only people make them complicated. In any event, the following table presents a hypothetical listing of combined frequency of judges' "yeses". It should be noted that the magnitude of the entries would be much greater in actual experimental practice.

SUBJECT	DEODORANT		
	A	B	C
1	3	2	1
2	2	2	0
3	2	2	1
4	3	2	1
5	3	3	2
6	2	3	1
7	3	2	0
8	3	3	2
9	3	3	2
10	1	2	0

First of all, the usual individual variations occur, but the main point is the consistently higher values for products A and B over C. (Note that "averages" are not needed and not presented.) In each comparison (A-C and B-C), perfect consistency is achieved in this hypothetical case. Ten out of ten events by the binomial yields a P of less than 1 in 1000. A comparison of A with B yields roughly a 50-50 split. Thus, product C is the "effective" deodorant of the three, remembering that the numbers represent the frequency of "can smells" by the judges.

The classical anova procedure for correlated, self-control data such as these adds one arithmetical manipulation. Besides considering and computing variation across columns (treatment effects), the correlation is

taken into account by dealing with variations across rows. If the correlation is substantial, this variation will be large, and when partialled out of the error variance, will leave the latter small thus enhancing the significance level. If the correlation is less than substantial, the investigator may have wasted his time in matching and in computing the correlational variation. It would be wise to inspect the data first.

In the case of the Whiff Test, the numbers are small and the spread so restricted that it hardly seems sensible to talk about correlation. Thus classical double classification anova for correlated data hardly seems applicable. The self-control design, however, paid off in that the simple binomial procedure allowed for rapid support of the inspectional analysis, namely, product C separated off for the judges from A and B.

To stamp in the point about correlated anova, there follow some data from a drive experiment where four pigeons were exposed to aperiodically reinforced responding at three different percentages of satiated body weight. The precautionary controls were, of course, exercised of using different orders of presentation of drive levels for each bird, measuring several times at each level, stabilizing body weight before measurement and so forth. The numbers represent responses in 30 minutes divided by 100 and rounded for simplification.

	DRIVE LEVEL		
S	75%	85%	95%
1	17	15	10
2	13	4	6
3	24	17	11
4	9	6	2

Several items are immediately obvious in this table. Across the wide range of responding represented, all birds show a diminution in frequency of response as drive is decreased. There is only one small reversal and a suggestion of approach to an asymptote appears. Many things could be done to these data statistically; little need be. Considerable correlation emerges in the data: birds starting high, stay high and vice versa. Double classification anova, teasing out the effects of drive (columns), self-control or correlation (rows) and error (remainder), yields significance supporting the obvious nature of the numbers.

The computational steps for the traditional double classification anova for correlated data are presented in detail in standard textbooks and need not be spelled out here. It seems worthwhile, however, to refer back to the rank procedure for the correlated data set-up that was presented in detail in the original Q & D manuscript. The Friedman Rank Anova is quite straightforward. Table 15 presents some data appropriate to it from an experiment on Thorndike's "spread of effect" but without reward or learning (Sheffield, 1949, Sheffield and Jenkins, 1952). College students simply wrote down several hundred numbers from 1 to 10 "chance" repetitions were lined up on the answer sheets and the percentage of repetition following these chance repeats was calculated.

In the Friedman procedure, the ranking takes place S by S across rows. The ranks are then summed by columns and substituted in a formula that yields a Chi Square - like number. The question being asked is whether the sum of ranks by columns pull far enough apart to warrant rejection of the Null hypothesis where the correlation (and the design

Table 15

Percent repetition in a "spread of effect" set-up without reward or learning.

S	CHANCE REPEATS	POSITION AFTER CHANCE REPEAT 1	2	3
1	299	26.1	11.4	11.7
2	303	17.5	15.8	13.1
3	276	20.3	10.2	10.6
4	289	19.4	11.7	10.0
5	318	20.4	13.5	11.9
6	318	24.5	13.2	13.2

matching) is considered by ranking across rows. It should be obvious that if every S is, say, highest under a particular condition, the sum of ranks for that condition will diverge from the others. Again, outlying scores are corrected for, at least in part.

Perusal of Table 15 indicates a clear sloughing off of Position 1 from behavior at the other two positions. (Note that chance in writing down the numbers 1 to 10 is 10% and that behavior at Position 2 exceeds this value by roughly a factor of two.) In all six cases percent repetition is higher at Position 1 than at either Positions 2 or 3 from a chance repeat. The binomial gives a probability of 1/64 for these two sets of events. The rank analysis of variance for correlated data yields a result consistent with the binomial scanning analysis, namely, a Chi Square of 9.1 and a P-value of .01.

Over and beyond any manipulation of the numbers, the important finding in this experiment is the occurrence of the "spread-of-effect" phenomenon in a setting where neither reward nor learning was operating. Since Thorndike labelled his original paper on the "spread-of-effect","A proof of the law of effect", data such as these "disprove" his proof and cast deep doubts on the formulation of the law of effect. As usual, a far simpler contiguity principle was operating to generate the findings, namely the number guessing habit sequences that Ss bring to the experimental situation so that when one number is anchored (in this instance on a chance basis), the several numbers associated with it in sequence follow. Evidence against the Law of Effect has been accumulating since before its inception. This type of result adds to the pile.

Classical anova could be applied to these data, but it hardly seems worth the effort in the light of the outcomes of the easier analyses. It must yield a significant result in view of the orderliness of the findings.

2. Repeated Measurements. It is a very common occurrence in behavioral research for the reactions of an organism to be recorded over a series of trials or in several blocks of time. For instance, the extinction curve deriving from the behavior of a rat in a Skinner box may well be divided into time portions. Or the latency or running time of a rat in a runway may be plotted trial-by-trial. Typically, in these situations an experimental treatment is applied to one or more groups and a control treatment to others with repeated measurements being taken for both groups. We are interested in the action of our experimental treatment, changes in behavior over time or trials and the interaction of the two, that is, systematic, differential changes in one group as contrasted with the other as time or trials go on. Certain experimental operations may contribute to the retardation or facilitation of acquisition or extinction. The effects emerge as we contrast an experimental with a control group over a series of trials or blocks of time. In extinction, for instance, a given procedure may result in retardation of the last half of extinction with little or no impact on behavior in the first half of extinction. This section is concerned with these trend matters. It might be noted that the "repeated measurement" set-up is essentially an extension of double classification anova. The same Ss are repeatedly tested, some under one set of conditions and other Ss under other condi-

tions, so that a correlational component is involved within each of the conditions.

At this juncture the reader should again be cautioned that the criterion for assessing data is adamant: If one can't see the effect in the numbers, it's very likely not there.

To start with a complex example and work back to the simple, Table 16 contains some results from a generalization-drive experiment with pigeons (Jenkins et al, 1958). After stabilization of responding on an APR schedule with one group at 90% of satiated body weight and the other at 70%, the size of the illuminated spot on the pecking window was varied systematically during brief extinction-generalization tests. These were repeated a number of times. Stabilized responding was used as the baseline to convert test responses to percentages to cut back on individual variability. The bird-by-bird data are contained in Table 16.

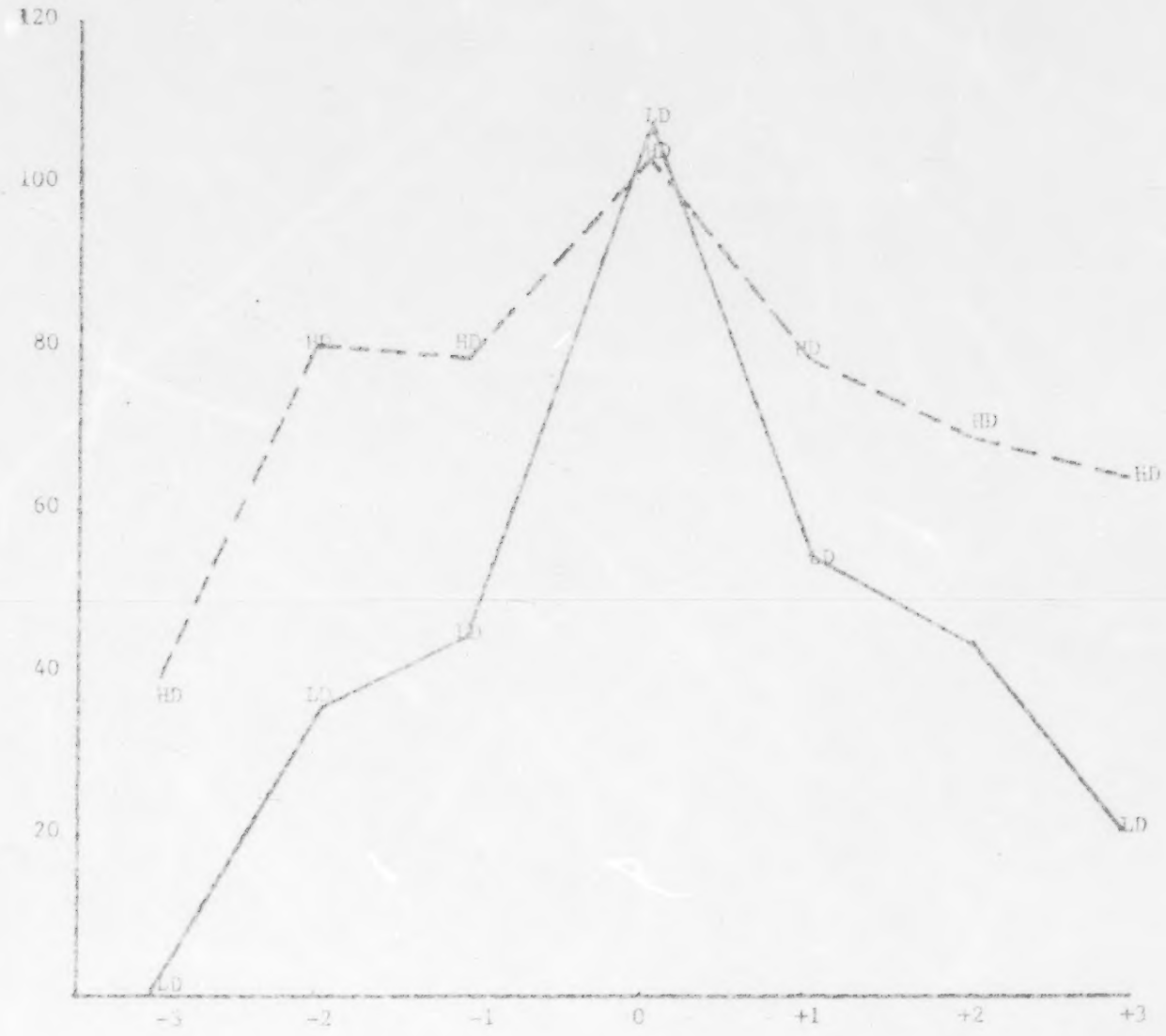
One could whip this series of numbers to a pulp statistically and squeeze nothing more from them than meets the naked eye. First things first: the individual generalization functions of the two birds nearest the median of their respective groups in training are plotted graphically in Fig. 1. From these two representations and without recourse to any statistical manipulation, it is obvious that several differential behavioral events have occurred. First, except for a couple of minor reversals, all birds exhibited consistent generalization decrement functions: as stimulus dissimilarity from the standard increased, responding decreased - the usual finding in this setting. Next, drive had an appreciable effect on responding with appreciably higher percentages appearing for the high-

Table 16

Generalization as a function of drive level in percentage terms.

(Jenkins et al, 1958)

	BIRD	SPOT SIZE						
		-3	-2	-1	0	+1	+2	+3
LOW DRIVE	1	01	34	47	106	55	43	20
	3	01	52	16	109	32	20	10
	4	18	26	67	114	51	65	39
	5	01	08	53	85	53	21	55
	6	02	24	36	72	46	21	16
	7	23	46	75	104	53	48	51
	8	54	38	98	115	58	41	32
Mean		15	32	56	100	49	37	32
HIGH DRIVE	9	23	85	103	90	56	57	47
	10	12	80	108	107	80	28	33
	11	55	66	113	110	89	69	74
	12	48	80	66	111	82	51	74
	13	41	80	79	104	79	69	64
	14	13	21	34	90	79	34	19
	15	44	71	104	112	117	85	80
	16	53	99	91	105	83	80	71
Mean		36	36	87	104	83	59	58



S P O T S I Z E

LD	#1	1	34	47	106	55	43	20
HD	#2	41	80	79	104	79	69	64

FIGURE 1 -

CENTRALIZATION AND DEVIATION

drive (HD, 70%) group than for the low-drive (LD, 90%) condition. Again, this is a common finding that, in an already conditioned piece of behavior, greater deprivation generates increased responding. The third, and most important effect, is the interaction of drive and stimulus change. Interaction means, of course, simply that behavioral changes associated with one experimental dimension of variation vary differentially with the application of some other treatment. Behavior changes as a joint function of the two dimensions, say, experimental and control. Exactly that happened here. As stimulus dissimilarity increases, the two generalization decrement functions pull apart with the HD group showing a flattening out and much less decrement while the LD continues to drop off with increased dissimilarity.

Across the board, careful study of Table 16 and Fig. 1 clearly support these inferences. Journal editors, however, require more elegant statistical manipulation. If these are applied, the three sources of behavioral variation turn out significant: drive, spot size and the interaction of the two. Such elaborate trend procedures may satisfy editors and those who are compulsive about their statistical analysis, but they can be frustrating and time-consuming for the behavioral scientist who can see the effects clearly in the data and wants to get on about his experimental business. However, this presentation is a dialogue not a diatribe.

A nice example of an apparent contradiction between simple, inspectional-type statistics and a more elaborate, complicated procedure derives from the data of Table 17. The numbers are extinction responses

Table 17

Pigeons' extinction responses in 20 min. periods with massed and distributed extinction.

		20 MINUTE PERIODS			
	S	1	2	3	Total
DISTRIBUTED	1	48	35	16	99
	2	27	6	13	46
	3	92	29	10	131
	4	171	31	9	211
	5	57	43	18	118
	6	56	17	4	77
	7	67	24	7	98
	8	18	6	0	24
	9	4	10	0	14
	10	44	38	22	104
Mean		58.4	23.9	9.9	92.2
Median		52.0	26.5	9.5	98.5
MASSED	1	14	8	40	62
	2	105	23	38	166
	3	71	92	9	172
	4	131	123	28	282
	5	31	0	34	65
	6	46	1	26	73
	7	49	5	23	77
	8	105	48	62	215
	9	54	5	23	82
	10	110	16	22	148
Mean		71.6	32.1	30.5	134.2
Median		62.5	12.0	27.0	115.0

for two independent groups of pigeons conditioned on a 100% reinforcement schedule with the major treatment being distribution of practice. One group was conditioned and extinguished in a series of brief sessions while the other was exposed to continuous conditioning and extinction without a break. The theoretical reasoning behind this experiment is quite straightforward. One position regarding extinction is that it constitutes a passive, decay process. The contiguity view, on the other hand, holds that extinction is a form of learning where, under conditions of radical stimulus change (particularly after 100% reinforcement), a new habit is acquired in the presence of a major portion of the original stimulus compound. In the case of pigeons operating in a Skinner Box, the new habit consists of doing something (usually not recorded) other than pecking the illuminated window. Since distribution of practice facilitates learning, and if extinction is learning, the latter should be speeded up by distributing extinction trials or sessions. Thus a crisp counter-opposing test of the two views of extinction are provided by this type of experiment. (Virginia Sheffield (1950) performed the classical investigation in this area.)

The numbers contained in Table 17 are a little complicated, but the trends are clear from inspection. All 10 birds in each condition show the decrement in behavior associated with extinction operations. There is considerable intra- and inter-group variability, but the data suggest a clear trend in the direction of the hypothesis of more rapid extinction for the group treated with distributed extinction. Or in other words, this group shows faster acquisition of some habit other than pecking the

window - whatever it may be. Furthermore, it seems as if the groups start quite close together early in extinction and pull apart in the last 20 minute period. As a matter of fact there is only one case in the massed group whose responses get onto the distributed distribution in the last 20 monutes of extinction. (It is noteworthy that 8 of the 10 massed extinction birds increase responding from the second to the third 20 minute period.) From these not so casual inspections, it would then appear that a case may be made for the significant action of 1) distribution of extinction practice, 2) extinction sessions and 3) interaction between the two with the functions pulling apart over sessions.

Classical statistics do not agree with these interpretations. The traditional trend analysis for repeated measurements shows that only extinction per se is significant, a point that is obvious in that all 20 birds showed decremental effects over sessions.

These contradictions need to be resolved. If one accepts the conclusions available from the classical analysis, a good deal of information is overlooked and the findings are equivocal with regard to the hypothesis entertained at the outset. It seems wasteful to follow this procedure and disregard some striking trends in the data. As an initial probe, the overall repeated measurement analysis may be useful, but it appears quite insensitive to the actual behavioral changes occurring. Thus we must resort to other techniques if we are to salvage a test of the hypothesis - and it is seems a highly worthwhile step. Several things may be done to the data. For one thing, conversion of the raw numbers to ranks helps a little in cutting back on the appreciable variability, but even

under these conditions, usual statistical significance is not achieved for the basic treatment variable of distribution of extinction practice. Another way to go is simply to analyze the data by fragments - even in the teeth of the objections that can be raised to piecemeal statistical treatment. For instance, a classical t-test applied to the extinction responses in the last 20 minute period yields a highly significant P-value, as it must from the almost non-overlapping nature of the distributions. But this procedure still leaves the situation somewhat openended. It could be argued, for example, that the distributed group (for whatever (chance) reason) started lower (but not significantly) in performance in extinction and ended up lower simply because of the built-in behavioral correlation. This is a possibility that must be considered. The obvious procedure is to convert extinction responses in the last 20 minutes to percentages, bird-by-bird, of the first 20 minutes of extinction. The median decrement in the distributed group was about 90% while in the massed group, it was only 49%. The corresponding means were 80% and 20%. The variability in the percentages was quite large so that a Mann-Whitney-Wilcoxon Rank T-test was employed. It yielded a P-value of .01. Splitting the percentages on the grand mean, the P-value associated with the Fisher-Yates Exact Test was .007 with a Phi Coefficient of .50.

All these additional rather detailed analyses in support of inspection prove out what one sees in the data, namely, a large and significant difference in the third 20 minutes of extinction with the distributed birds losing the old behavior and acquiring the new more rapidly than the birds exposed to the massed extinction procedure. Since differences

were small and insignificant in the first 20 minutes of extinction, this significant finding clearly indicates a pulling apart of the behaviors as a joint function of distribution of practice and extinction sessions, i.e., suggests a clearcut interaction effect. It hardly seems necessary to go farther with statistics in this instance. It does, however, provide a nice example of a basic caution: caveat emptor when the gifts are those of traditional statistical analysis applied to behavioral data with all its vagaries and complications. Do not buy a statistical pig in a poke.

To end this section with a relatively simple example, Table 18 was constructed. It consists of extinction responses - coded, rounded and simplified - from an experiment on reinforcement theory in which a brief flash of light was thrown on the pecking window during pigeons' extinction as a substitute for the presentation of food during prior aperiodically reinforced responding. This increase in stimulation (as well as change) should serve as a reinforcing agent by the contiguity position and contrary to the drive-reduction view. Its reinforcing property lies in its ability to change behavior, to bring about momentary pauses as changes in the pigeons' behavior and thereby maintain the behavior above the extinction level of a control group without this light-up treatment.

The upshot of Table 18 is straightforward. Behavior was maintained by the light-up although decremental extinction effects appeared in the behavior of both groups. The behavior started at about the same level in the first half of extinction and pulled apart in the second half. Six out of six birds showed decrement in behavior (P of .016 by the binomial) and the three light-up birds exceeded the three controls in the second half

Table 18

Pigeon extinction responses coded and rounded, with and without brief periods of increased illumination substituted for food.

	S	1st Half	2nd Half
Light-Up.	1	8	5
	2	7	4
	3	6	3
No Light-Up	1	8	2
	2	7	1
	3	5	1

of extinction (P of .05 by the Arrangement Technique). It follows that, with two groups starting at the same point and ending up at different points with significance accruing to time and treatment differences, interaction between the two dimensions is also significant. The proof lies not in the classical statistical pudding, but it may be employed as a supplementary procedure. All conclusions from the "rough and ready" analyses are supported by the more traditional approach: significance emerges for the three sources of experimental treatment, extinction over sessions and the interaction of the two. It helps when the classical, far more cumbersome procedure generates results consistent with the quick and dirty ones. Inspection again pays off.

3. "Simple" Factorial Design. "Simple" factorial design is fairly straightforward; "simple" factorial analysis of the outcome of the design is far from simple. In the former instance, factorial design involves the "simultaneous" application of two dimensions of experimental variation. It is clearly not "simultaneous" because independent groups of Ss are involved. In the simplest case there are two experimental treatments with just two conditions each, making up a two-by-two layout of four cells in all. The generalized case or prototype is this:

VARIABLE A

VARIABLE B	CONDITION 1	CONDITION 2
CONDITION 1		
CONDITION 2		

Thus the four independent groups constituting the cells of this set-up are: A1, A2, B1 and B2. How many cases are treated in each cell and other considerations are a function of the nature of the problem, the experimental treatments, the behavioral measurements involved, and the like. Analytically, the procedure consists of teasing out the effects of Variable A separately, those of Variable B by itself without regard to A and finally the joint action or interaction of the two dimensions of variation simultaneously, namely, the diagonal cells A1B1 plus A2B2 versus A1B2 plus A2B1. Probably the simplest paradigm for remembering the factorial lay-out is a stimulus change or generalization experiment where one dimension is degree of stimulus dissimilarity from the originally conditioned stimulus and the other experimental treatment constitutes degrees of drive, partial reinforcement, distribution of practice and so forth. The upper left hand cell combined with the lower right hand cell constitute the conditions of no stimulus change; the other diagonal cells are the ones treated with change. This point will be spelled out b ow.

This is as good a place as any to pinpoint the nature of the interaction source of variation in behavior particularly and in statistics secondarily. We will consider only the simple interaction case where there are two treatments and a single interaction. More complex interactions of three of more variables may be comprehensible to sophisticated mathematicians in a statistical sense, but their behavioral meaning appears to rapidly fade away.

"Interaction" is relatively easy to describe in behavioral terms. The question is: do the responses change differentially with the application

of both variables; are they a joint function of the two treatments? Put more simply, is behavior different under one set of experimental values than under another? Does behavioral change hinge on the combined action of the two experimental variations? Examples may help clarify the matter.

There are four major combinations of events with two treatments. (There are several others, but they are minor for our purposes.) They are:

A. Where only one of the experimental treatments has an impact on behavior;

B. Where both treatments influence behavior on a large scale leaving little behavioral variation left over for interaction effects;

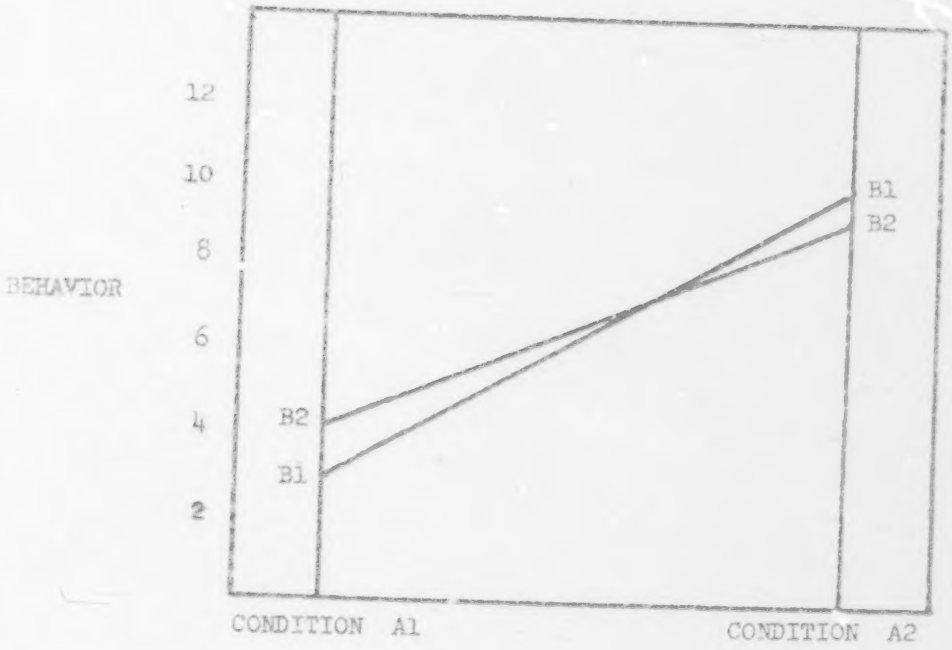
C. Where interaction accounts for most of the behavioral variance with little remainder for the two experimental treatments;

D. And where all three primary sources of variation - the two basic variables and the interaction - have a big impact on behavior.

Each of these cases will be considered in turn.

CASE A: The Operation of One Variable. The accompanying chart shows in numbers and graphically what happens in the hypothetical case where one value influences behavior in the two-by-two set-up and little impact is exerted by the other variable or by the joint action of the two variables (interaction). Here as values of Variable A increase, behavior increases regardless of whether condition B1 or B2 is involved. The two functions, so to speak, go up together. The marginal sums in the tabular material are the key to inspectional analysis. These reveal an increase by more than a factor of two as we go from Condition A1 to Condition A2. No difference emerges between Conditions B1 and B2 and little between the diagonal cells

CASE A: THE OPERATION OF ONE VARIABLE



VARIABLE A

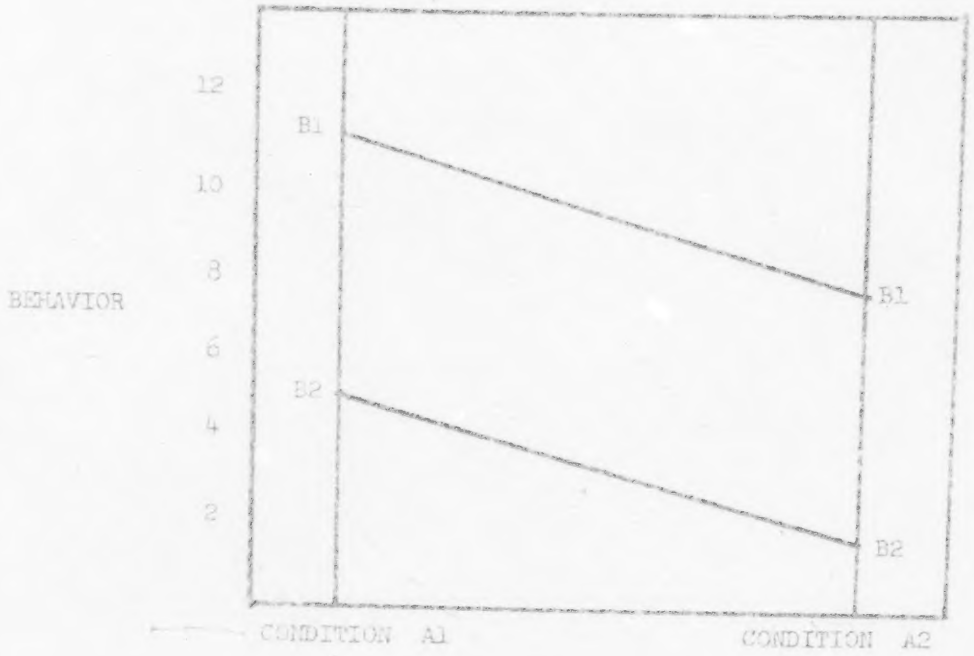
VARIABLE B	COND. 1	COND. 2	SUM
COND. 1	2 3 6	7 10 11	39
COND. 2	1 4 5	8 9 12	39
SUM	21	57	Diagonals: 40 vs 38

(A1B1 plus A2B2 versus A2B1 plus A1B2). Analysis of the data by a standard t-test would reveal a highly significant difference between Conditions A1 and A2 without regard to variable B. As a matter of fact the A1 and A2 distributions do not overlap. The more elaborate interactive analysis reveals essentially the same outcome. Significance accrues to Variable A but to neither Variable B nor the interaction of Variables A and B.

This kind of finding might emerge from a "perceptual" experiment in which "Levelers" and "Sharpeners", perceptually defined, were selected to constitute Variable B and Variable A consisted of success or failure in learning or problem solving. The success-failure dimension influences behavior, but not the perceptual variable.

CASE B: The Operation of Both Variables. The chart presents the data for the case where both variables have a large and significant impact on behavior to such an extent that little behavioral variation is left over for the interaction of the two dimensions of variation. Again, the effects are clear in the marginal totals with the situation rigged so that the diagonal cells end up with the same sums. From these marginals it is also apparent that Variable B has a larger behavioral effect than Variable A, but that both operate on an appreciable scale. Instances where this kind of finding emerges are fairly common in behavioral research. An example that immediately comes to mind is the experimental case where partial reinforcement and cue change are applied "simultaneously". In this hypothetical case we have two degrees of stimulus change (Variable A), a control condition of "no change" and one degree of fairly marked change. Variable B is the reinforcement schedule and, in a typical experiment of this

CASE B: THE OPERATION OF BOTH VARIABLES



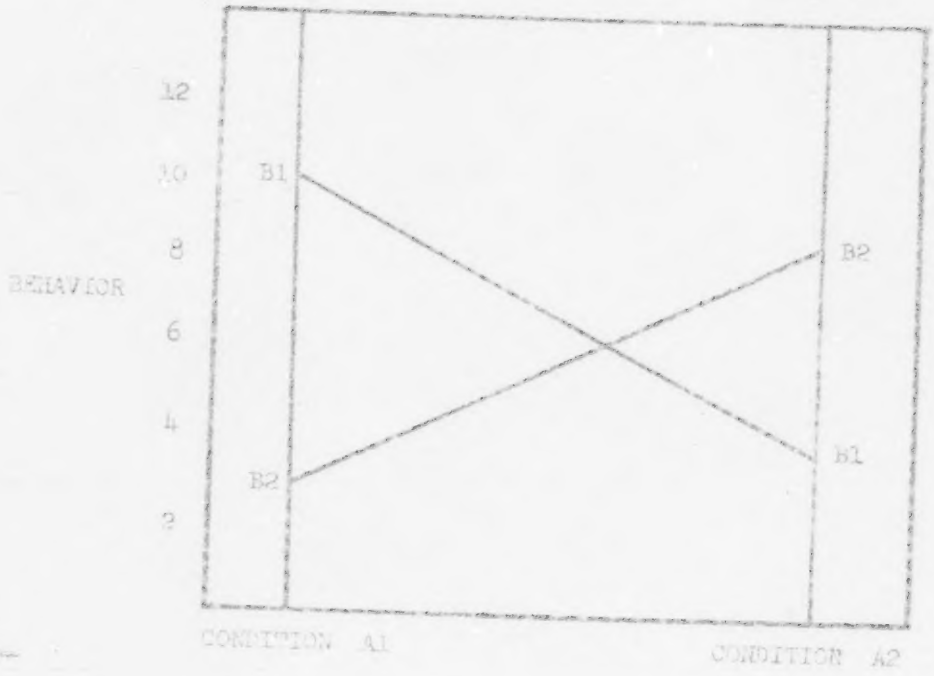
VARIABLE A

VARIABLE B	COND. 1	COND. 2	SUM
COND. 1	12 11 10	9 8 7	57
COND. 2	6 5 4	3 2 1	21
SUM	48	30	Diagonals: 39 vs 39

nature with human or infra-human organisms, the two conditions would most likely be 100% and 50% reinforcement. The hypothetical findings presented are not too far off the mark of actual findings in real-life experimental settings (viz Rickard, 1959). Of passing interest is the fact that a journal editor once turned down a paper containing this type of finding on the grounds that the interaction had to be significant. It's obvious, however, that if the two major sources of variatior have a "whopper" impact on behavior as in this instance, there cannot be much behavior left over for interaction. Possibly a formal academic course in inspectional analysis is called for. In any event, treatment of these data by any appropriate analysis supports what can be seen: the two variables have a significant influence on behavior. If one wished to analyze the data without recourse to the elaborate procedures, inspection reveals non-overlapping distributions for both sub-groups along both experimental dimensions. In all instances, three events exceeding three others yields a probability of .05 by the Arrangement Technique.

CASE C: The Operation of Interaction. There are some instances where behavioral change pivots on the joint action of two dimensions of variation. These are typically cases that have a behavioral impact when applied alone to one of the values of the other experimental treatment, but operate differentially when several values of the second variable are included. Such a hypothetical case is presented in the accompanying chart. The differential effects of Variable A on Variable B are immediately obvious. Behavior under Condition B1 decreases as A increases while under Condition B2 it increases with A. The numbers reflect this situation in

CASE C: THE OPERATION OF INTERACTION



VARIABLE A

	COND. 1	COND. 2	SUM
COND. 1	7 10 11	1 4 5	38
COND. 2	2 3 5	8 9 12	40
SUM	39	39	Diagonals: 57 vs 21

VARIABLE B

showing the marginal totals to be about the same by rows and columns, but to differ appreciable on the diagonal sums where the interaction effect operates. Behavior under Conditions B1 and B2 pull apart and, as a matter of fact fail to overlap, but in opposite directions for the two values A1 and A2. This is clear interaction.

Analysis of these data will show the interaction term to be highly significant while, across the board, neither Variable A nor B has a significant effect. There is no contradiction here. Taken alone A has a clear effect on B1 and taken alone it also has a clear effect on B2. But taken together the effects of A on both B1 and B2 cancels out. It seems clear that averaging the curves in the figure at the two sets of points will yield no change and zero slope. Given one value of A, behavior under B will differ, depending on whether the condition is B1 or B2. Behavior thus depends on both variables; to specify it one must know the values of both A and B. Behavior covaries jointly with the action of both dimensions of variation.

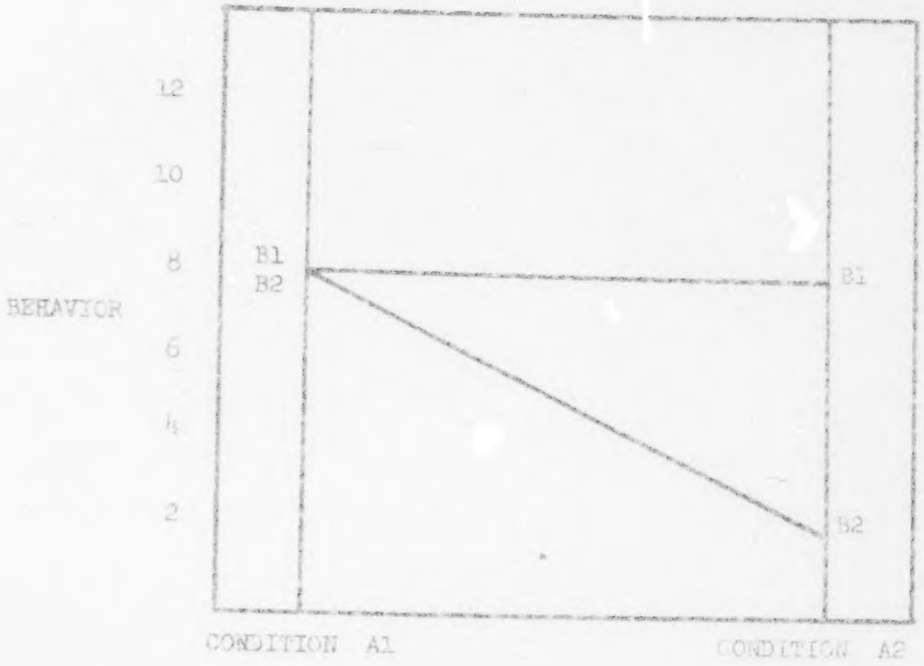
An actual experimental example may help stamp in the point. Findings such as those in Case C emerge in studies of rats' behavior in open fields. With drive (food deprivation) as a primary variable, behaviors classified as Gross Movements (GM consisting of locomotion and rearing responses) and Fine Movements (FM involving washing, grooming, scratching, sniffing and the like) operate quite differently. As drive increases, GM increase and FM decrease. In other words, the rats under high drive spend a good deal of their time running around and rearing up on their hind legs. With appreciably lower drive levels, FM increase markedly in frequency with

the rats sitting around grooming rather than locomoting or rearing. Thus in the accompanying representations, Variable A constitutes drive, while B2 consists of Gross Movements, and B1 of Fine Movements. The two sets of behaviors operate in a diametrically opposed direction as a function of the drive variable. As drive increases, one set of behavior increases while the other decreases. As drive decreases, the converse case holds.

CASE D: All Three Sources of Variation Operating. Finally there is the situation where both basic variables influence behavior along with their joint action. The accompanying tabular and graphical representations depict this state of affairs. It can be seen that B1 and B2 pull apart as one goes from A1 to A2. This is clearly the most striking feature of the representations. This differential reflects the decrement in behavior in Condition B2 as contrasted to the lack of change in B1 proceeding from A1 to A2. The situation is shown in the marginal totals where a clearcut differentail emerges for both Variables A and B as well as for the diagonals. Analysis of these numbers by the traditional procedures yields significance for all three sources of variation.

A case somewhat akin to this has already been cited in the repeated measurement section where the example was given of distribution of extinction generating retarded decremental effects for the massed group and/or facilitated decremental effects for the distributed extinction condition. It will be recalled that classical statistics did not uncover a significant interaction term that was visible in the data, but that a subcomparison of the latter part of extinction strongly supported a differential pulling apart of the massed and distributed curves and, thereby, an appreciable

CASE D: ALL THREE SOURCES OF VARIATION OPERATING



VARIABLE A

VARIABLE B		COND. 1	COND. 2	SUM
	COND. 1	9 8 7	5 8 7	48
	COND. 2	9 8 7	3 2 1	30
	SUM	48	30	Diagonals: 48 vs 30

interaction effect.

Another case in point is an experiment by Rowe (1955) in which a test of the Skaggs-Robinson hypothesis was conducted with the rearing response of rats in an open field. Hand removal reinforcement for rearing was employed and an extinction test run with total number of rears counted. Two basic conditions were involved: Generalization Decrement and Skaggs-Robinson. In both groups the rearing response was built in by removing the rats by hand when they were in a full rear in a particular open field (A). Two additional open fields were constructed differing in shape, height, illumination, texture and color of walls and floor, and the like, one (C) quite different from Field A and the other (Field B) judged to be midway between A and C. In one experiment dealing with Generalization Decrement (GD), the rats were conditioned in Field A and one-third tested in A, B and C. In the Skaggs-Robinson condition all rats were similarly trained in Field A and, in addition, one-third were given additional hand reinforced training in Fields A, B and C. All the latter rats were then returned to Field A for their free-responding extinction test.

Theory and previous data clearly suggest straight decremental effects for the GD rats as dissimilarity increases from Field A through B to C. The Skaggs-Robinson hypothesis suggests that, as dissimilarity of interpolated learning increases, interference effects increase at first and then decrease as dissimilarity becomes maximal. In other words, additional training on the same material contributes to over-learning. Learning of somewhat different materials interferes with retention of the originally learned responses; and when dissimilarity is at the limit, little or no

interference with original retention occurs since the two sets of stimuli and responses have little to do with one another. Thus the prediction from this position is for a U-shaped function with retention maximal (and interference minimal) at the extremes of the dissimilarity continuum and retention interferred with the most in the middle range of dissimilarity. Further, the U-shaped function should be asymmetrical with continued practice on the originally learned materials producing maximum retention above the level attained with interpolated practice on quite dissimilar materials at the other extreme.

The results, given below, support both the GD and Skaggs-Robinson hypotheses. The numbers represent rearing responses of the median rat in each group in a 10 minute extinction test period.

GENERALIZATION DECREMENT

AA	AB	AC
69	64	58

SKAGGS-ROBINSON

AAA	ABA	ACA
77	50	67

Both the GD and Skaggs-Robinson functions emerge clearly in these data. The GD finding declines in an orderly fashion while an asymmetrical parabola appears in the Skaggs-Robinson data. Replication supported these findings (Rowe, 1955).

Inspection indicates essentially no difference between the two con-

ditions where AA and AAA are compared. At the other two points (AB vs ABA and AC vs ACA) significant differences are suggested. It is to be noted that they are in opposite directions with the GD higher at the midpoint and the Skaggs-Robinson higher in the extreme change condition. In passing it might be noted that there was practically no overlap between the AA and AC performance in the GD instance and between AAA and ACA on the one hand and ACA on the other in the Skaggs-Robinson case.

A fair amount of variability characterized these data and the overall factorial analysis merely suggests significance for the two primary sources of variation and their interaction. In the final replicated findings a higher level of significance was achieved. At the least the findings are highly suggestive and indicate what behavioral changes can be achieved with small Ns and variables with large experimental effects. They further underscore the need for inspect onal analysis and non-necessity of elaborate statistical analysis.

4. Experimental Examples of Factorial Design. A couple of actual examples from the laboratory may help tie down these several complicated points concerning interaction. Table 19 contains some data from the performance of rats in an open field. Half the rats were given one-trial conditioning with hand-removal reinforcement in a field with cues minimized (small number of cues, S) and the other half the same treatment in a field with a large number of cues (L). Half of each group was then given a one-trial test in the same field and the other half in the different field. Latency of the rearing response was the index of behavior and the entries in Table 19 represent difference scores between latency of the

Table 19

Difference scores in sec. between one training and one test trial for hand removal reinforcement of the rearing responses in four groups of rats exposed to a small (S) number of cues or large (L).

CUES ON TEST		CUES IN TRAINING: Small	CUES IN TRAINING: Large
		4	1
		12	1
	Small	11	2
		9	0
		5	3
	Median	9	1
		1	10
		2	4
	Large	6	6
		0	5
		3	7
	Median	2	6